

An Introduction to Real-Time Operating Systems and Schedulability Analysis

Marco Di Natale
Scuola Superiore S. Anna

Outline

- Background on Operating Systems
- An Introduction to RT Systems
- Model-based development of Embedded RT systems
 - the RTOS in the platform-based design
- Scheduling and Resource Management
- Schedulability Analysis and Priority Inversion
 - The Mars Pathfinder case
- Implementation issues and standards
 - OSEK

Credits

- Paolo Gai (Evidence S.r.l.) – slides on EDF and OSEK
- Giuseppe Lipari (Scuola Superiore S. Anna) – slides on OS
- Manas Saksena (TimeSys) – examples on blocking time comput.
- From Mathworks Simulink and RTW manuals – slides on RT blocks

Background on Operating Systems

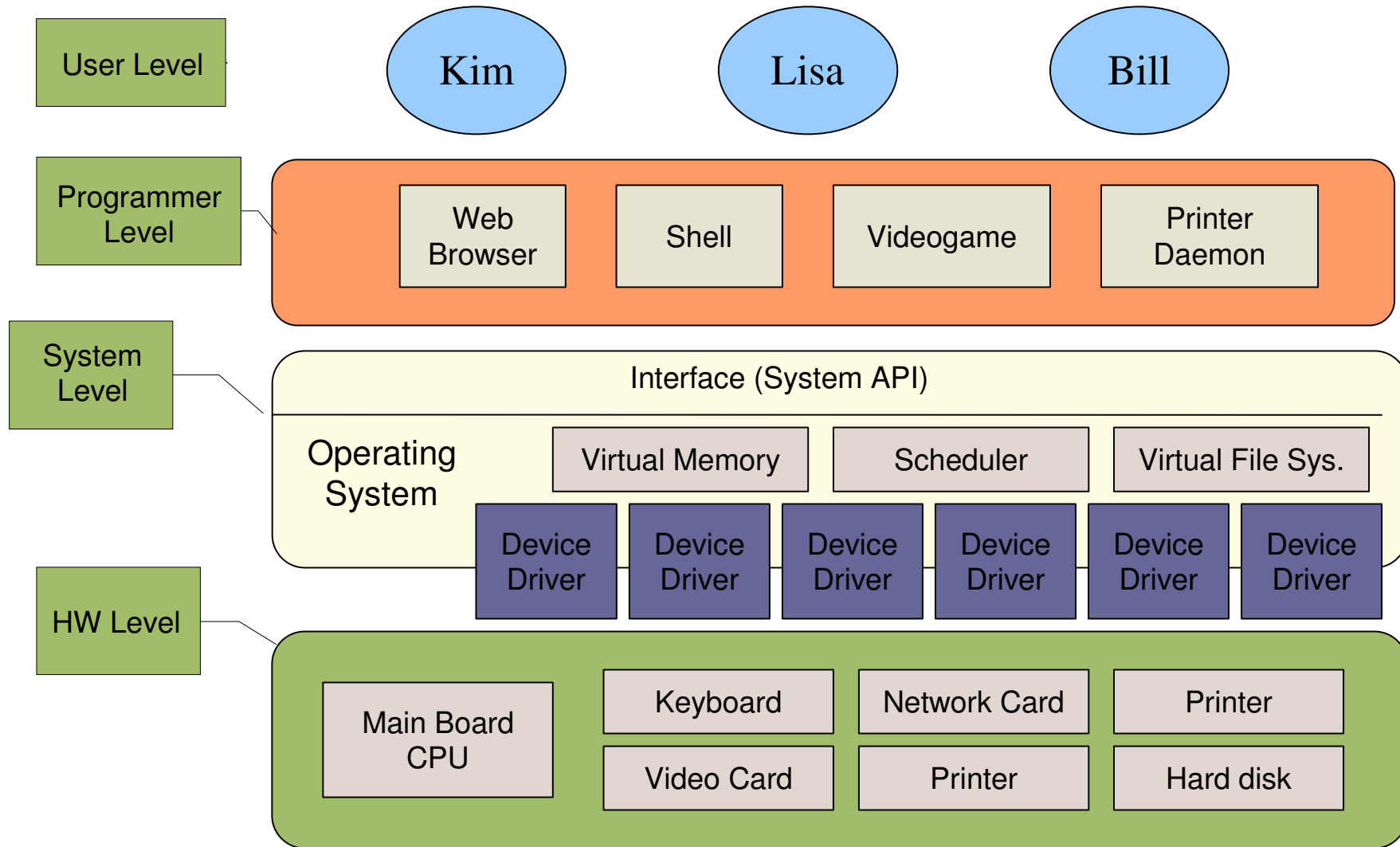
Fundamentals

- **Algorithm:**
 - It is the logical procedure to solve a certain problem
 - It is informally specified as a sequence of elementary *steps* that an “execution machine” must follow to solve the problem
 - It is not necessarily expressed in a formal programming language!
- **Program:**
 - It is the implementation of an algorithm in a programming language
 - Can be executed several times with different inputs
- **Process:**
 - An instance of a program that given a sequence of inputs produces a set of outputs

Operating System

- An operating system is a program that
 - Provides an “*abstraction*” of the physical machine
 - Provides a simple interface to the machine
 - Each part of the interface is a “*service*”
- An OS is also a resource manager
 - The OS provides access to the physical resources of a computing machine
 - The OS provides *abstract resources* (for example, a file, a virtual page in memory, etc.)

Levels of abstraction



Abstraction mechanisms

- Why abstraction?
 - Programming the HW directly has several drawbacks
 - It is difficult and error-prone
 - It is not portable
 - Suppose you want to write a program that reads a text file from disk and outputs it on the screen
 - Without a proper interface it is virtually impossible!

Abstraction Mechanisms

- Application programming interface (API)
 - Provides a convenient and uniform way to access to one service so that
 - HW details are hidden to the high level programmer
 - One application does not depend on the HW
 - The programmer can concentrate on higher level tasks
 - Example
 - For reading a file, linux and many other unix OS provide the **open()**, **read()** system calls that, given a “file name” allow to load the data from an external support

the need for concurrency

- there are many **reason for concurrency**
 - functional
 - performance
 - expressive power
- **functional**
 - **many users** may be connected to the same system at the same time
 - each user can have its own processes that execute concurrently with the processes of the other users
 - perform **many operations** concurrently
 - for example, listen to music, write with a word processor, burn a CD, etc...
 - they are all different and independent activities
 - they can be done “at the same time”

the need for concurrency (2)

- performance
 - take advantage of **blocking time**
 - while some thread waits for a blocking condition, another thread performs another operation
 - parallelism in **multi-processor machines**
 - if we have a multi-processor machine, independent activities can be carried out on different processors at the same time
- **expressive power**
 - many control applications are inherently concurrent
 - concurrency support helps in expressing concurrency, making application development simpler

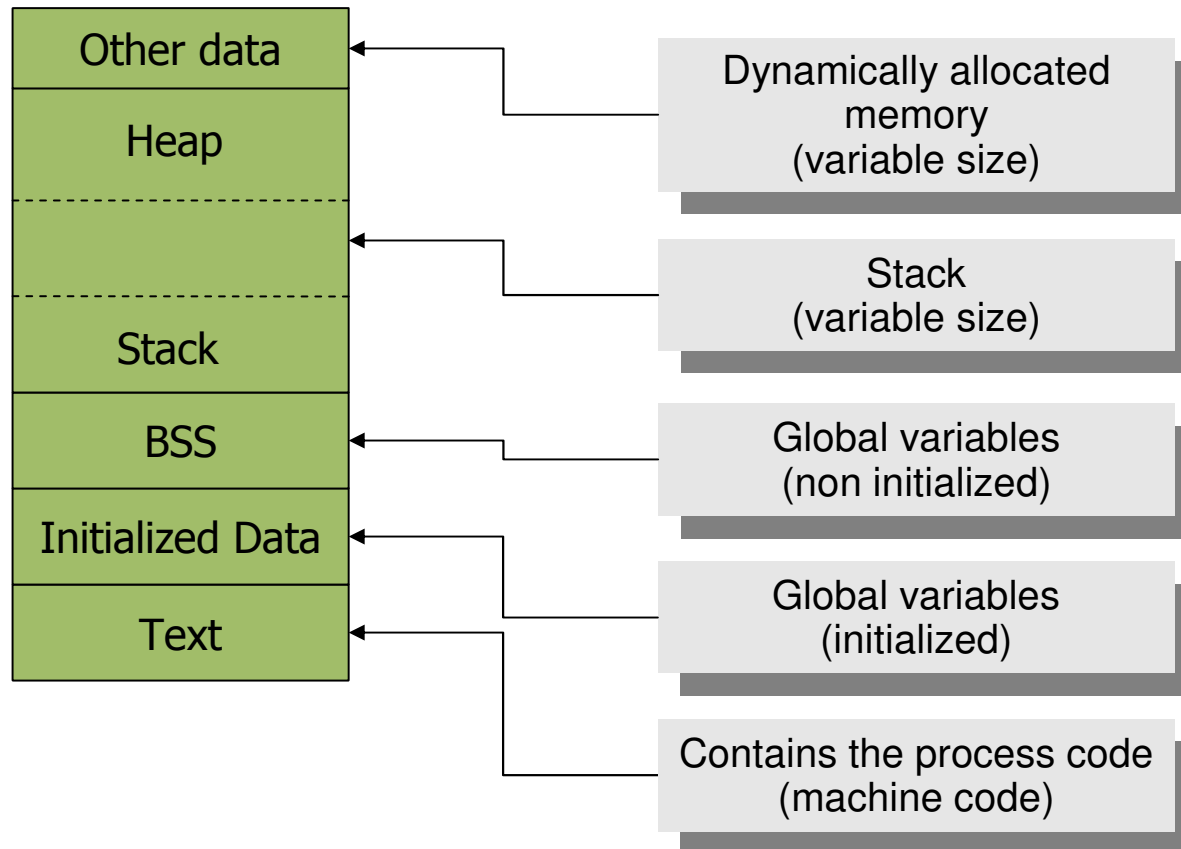
theoretical model

- a system is a set of **concurrent activities**
 - they can be processes or threads
- they **interact** in two ways
 - they **access the hardware resources**
 - processor
 - disk
 - memory, etc.
 - they **exchange data**
- these activities **compete** for the resources and/or **cooperate** for some common objective

Process

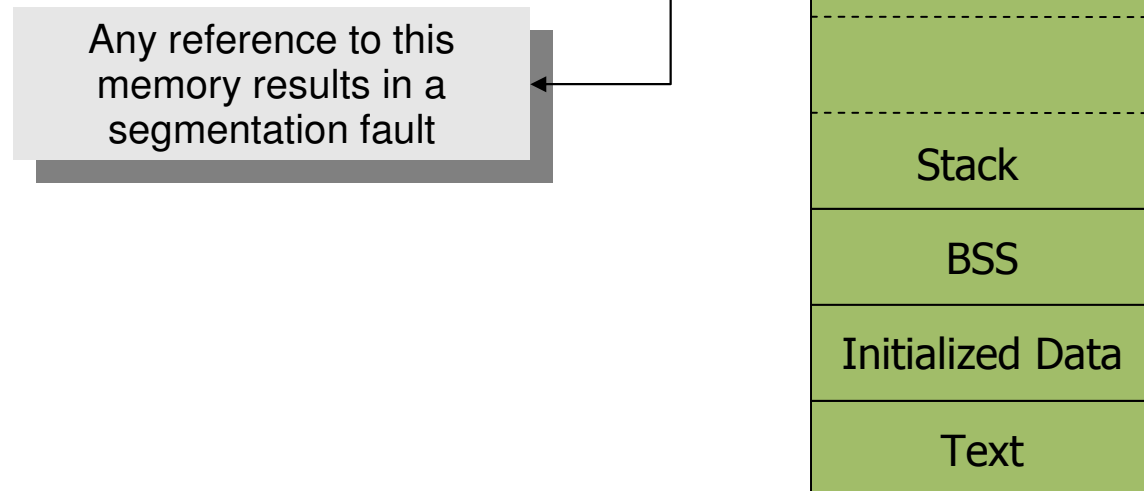
- The fundamental concept in any operating system is the “process”
 - A process is an executing program
 - An OS can execute many processes at the same time (concurrency)
 - Example: running a Text Editor and a Web Browser at the same time in the PC
- Processes have separate memory spaces
 - Each process is assigned a private memory space
 - One process is not allowed to read or write in the memory space of another process
 - If a process tries to access a memory location not in its space, an exception is raised (Segmentation fault), and the process is terminated
 - Two processes cannot directly share variables

Memory layout of a Process



Memory Protection

- By default, two processes cannot share their memory
 - If one process tries to access a memory location outside its space, a processor exception is raised (trap) and the process is terminated
 - The “Segmentation Fault” error!!



Processes

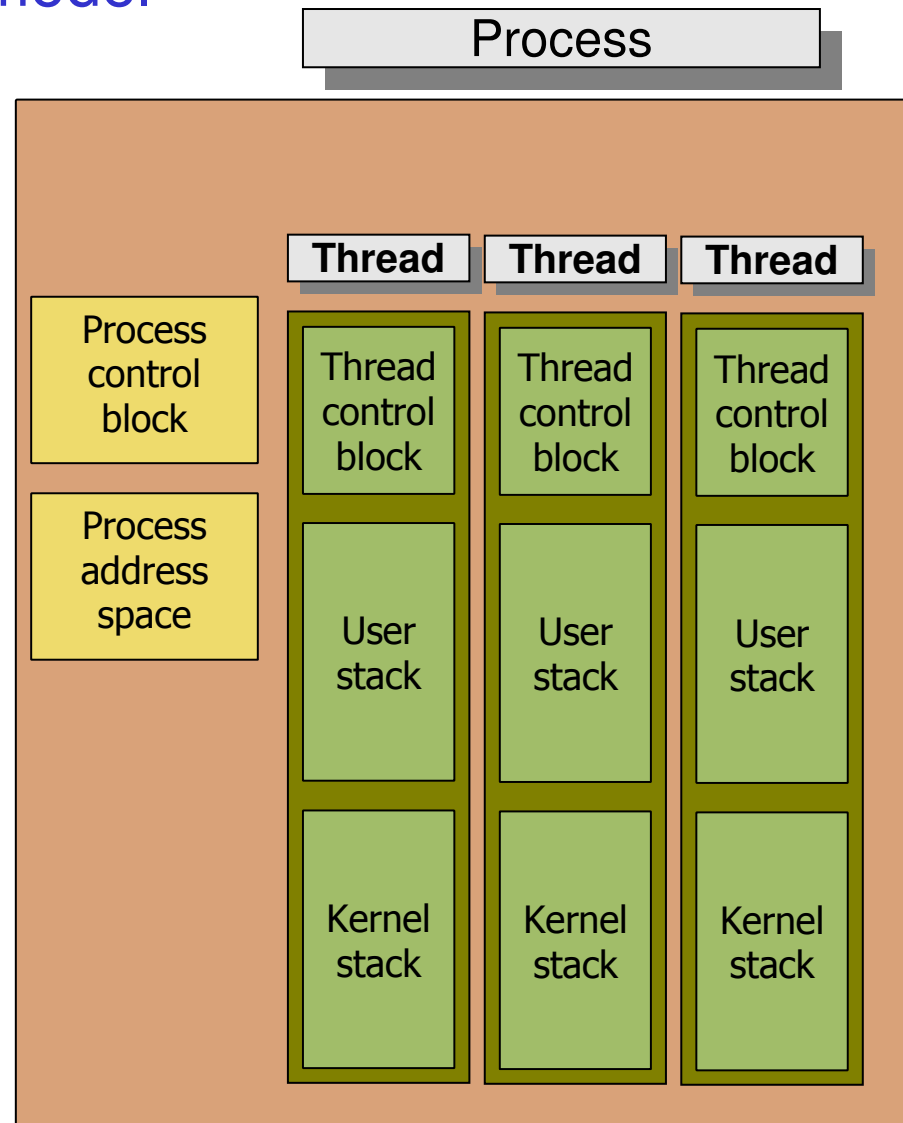
- We can distinguish two aspects in a process
- **Resource Ownership**
 - A process includes a virtual address space, a process image (code + data)
 - It is allocated a set of resources, like file descriptors, I/O channels, etc
- **Scheduling/Execution**
 - The execution of a process follows an execution path, and generates a trace (sequence of internal states)
 - It has a state (ready, Running, etc.)
 - And scheduling parameters (priority, time left in the round, etc.)

Multi-threading

- Many OS separate these aspects, by providing the concept of thread
- The process is the “resource owner”
- The thread is the “scheduling entity”
 - One process can consists of one or more *threads*
 - Threads are sometime called (improperly) lightweight processes
 - Therefore, on process can have many different (and concurrent) traces of execution!

Multi-threaded process model

- In the multi-threaded process model each process can have many threads
 - One address space
 - One PCB
 - Many stacks
 - Many TCB (Thread Control blocks)
 - The threads are scheduled directly by the global scheduler



Threads

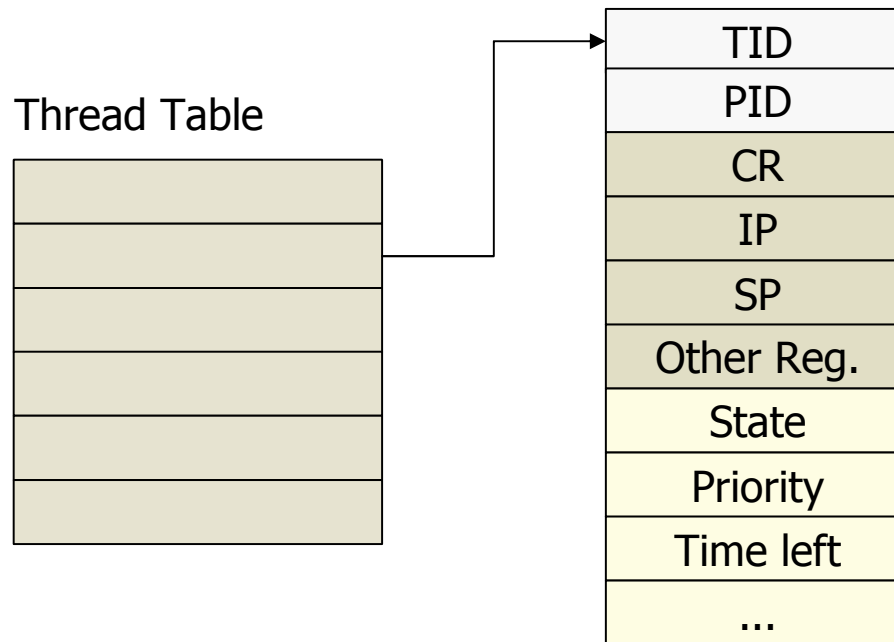
- Generally, processes do not share memory
 - To communicate between process, it is necessary to use OS primitives
 - Process switch is more complex because we have to change address space
- Two threads in the same process share the same address space
 - They can access the same variables in memory
 - Communication between threads is simpler
 - Thread switch has less overhead

Threads support in OS

- Different OS implement threads in different ways
 - Some OS supports directly only processes
 - Threads are implemented as “special processes”
 - Some OS supports only threads
 - Processes are threads’ groups
 - Some OS natively supports both concepts
 - For example Windows NT
- In Real-Time Operating Systems
 - Depending on the size and type of system we can have both threads and processes or only threads
 - For efficiency reasons, most RTOS only support
 - 1 process
 - Many threads inside the process
 - All threads share the same memory
 - Examples are RTAI, RT-Linux, Shark, some version of VxWorks, QNX, etc.

The thread control block

- In a OS that supports threads
 - Each thread is assigned a TCB (Thread Control Block)
 - The PCB holds mainly information about memory
 - The TCB holds information about the state of the thread

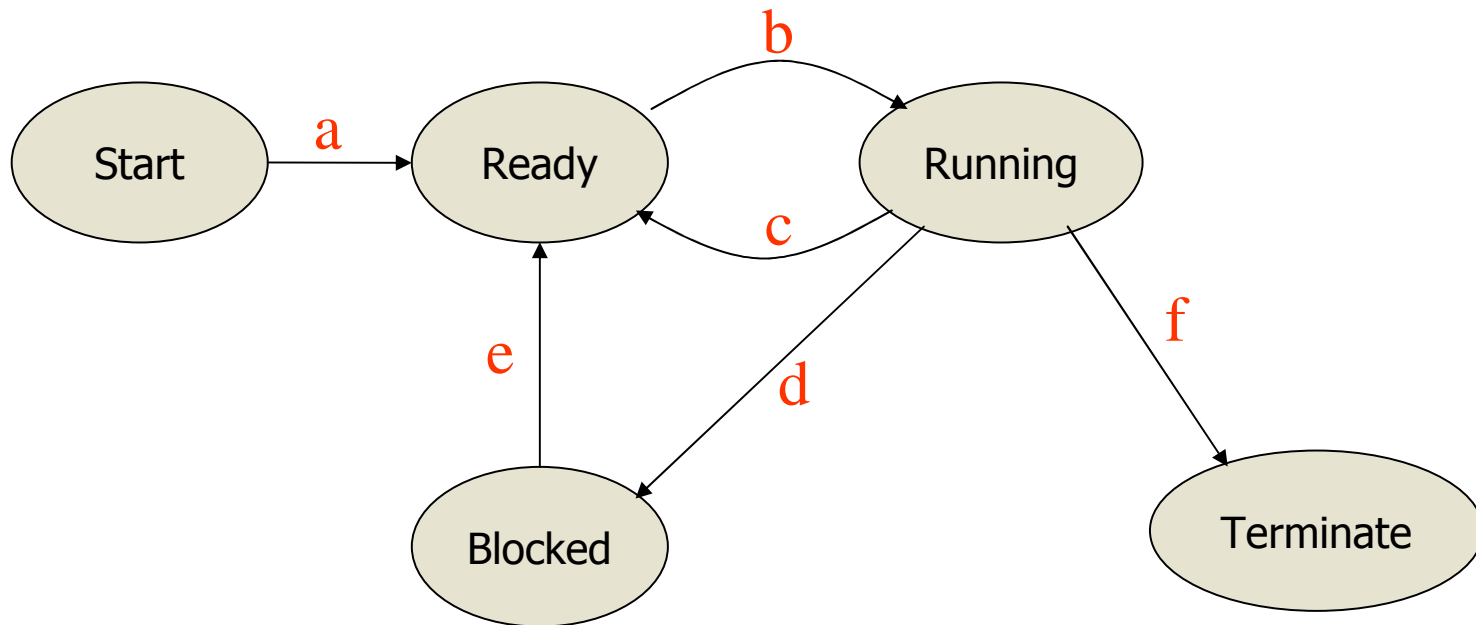


Thread states

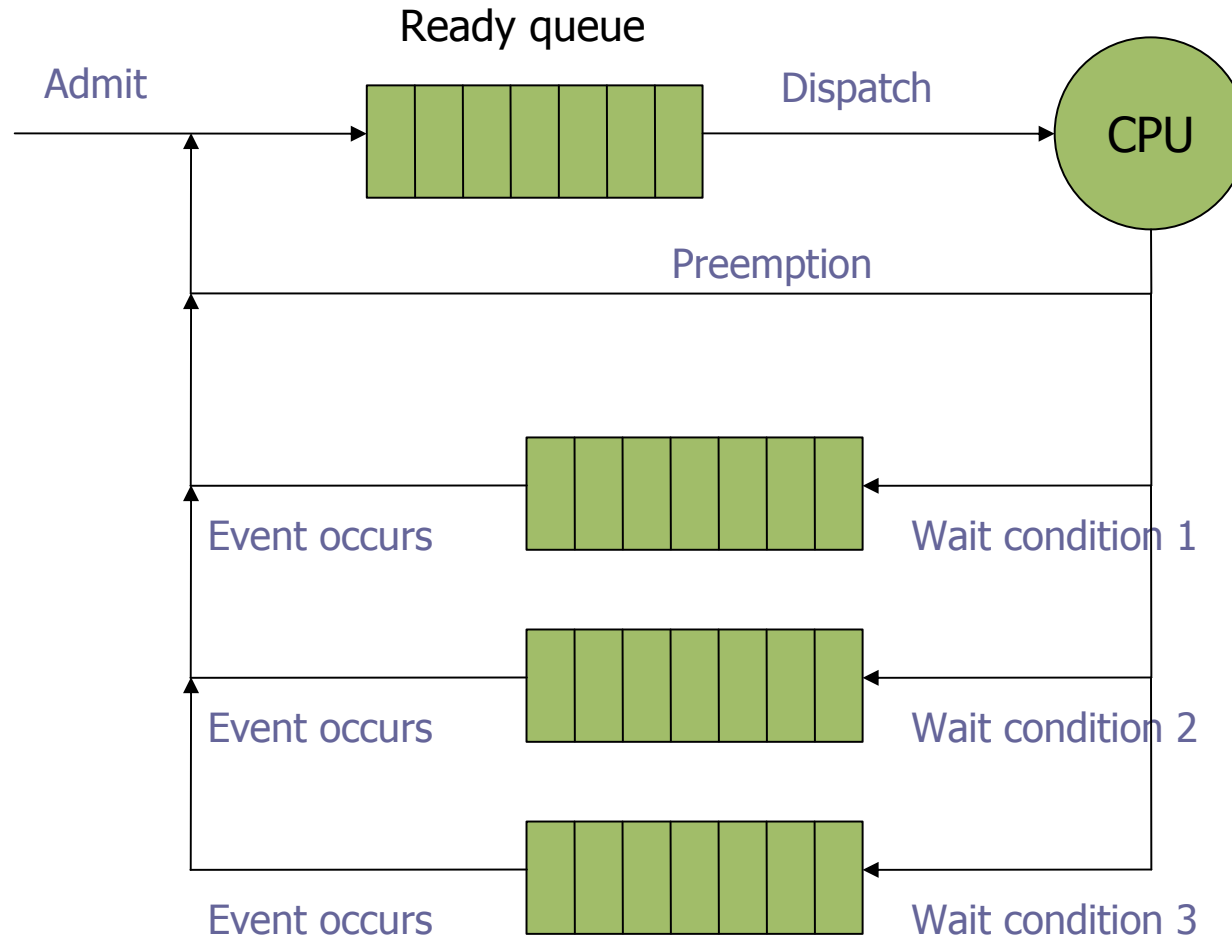
- The OS can execute many threads at the same time
- Each thread, during its lifetime can be in one of the following states
 - Starting (the thread is being created)
 - Ready (the thread is ready to be executed)
 - Executing (the thread is executing)
 - Blocked (the thread is waiting on a condition)
 - Terminating (the thread is about to terminate)

Thread states

- | | | |
|----|-------------------|--------------------------------------|
| a) | Creation | The thread is created |
| b) | Dispatch | The thread is selected to execute |
| c) | Preemption | The thread leaves the processor |
| d) | Wait on condition | The thread is blocked on a condition |
| e) | Condition true | The thread is unblocked |
| f) | Exit | The thread terminates |



Thread queues

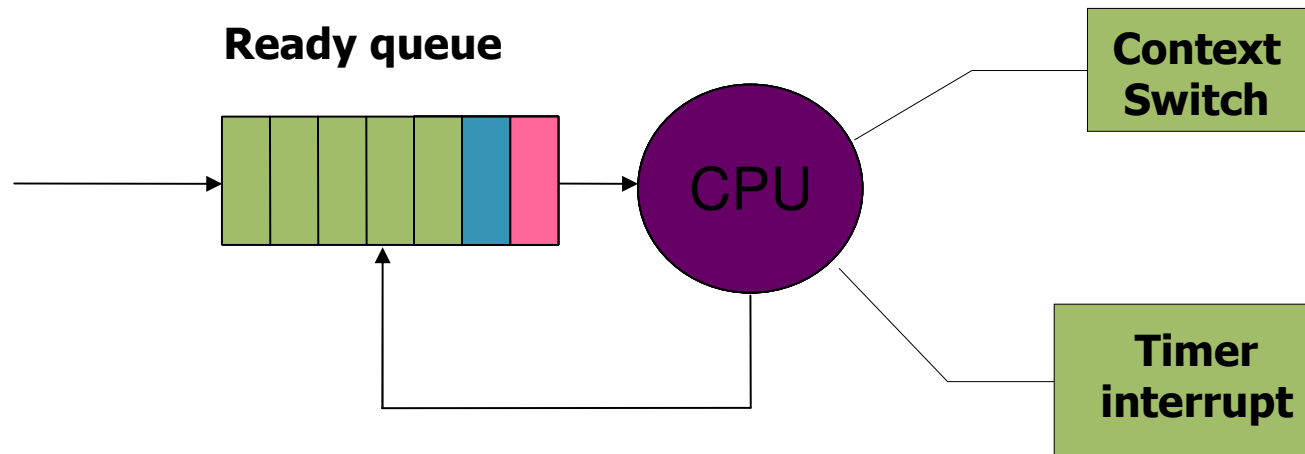


Context switch

- It happens when
 - The thread has been “preempted” by another higher priority thread
 - The thread blocks on some condition
 - In time-sharing systems, the thread has completed its “round” and it is the turn of some other thread
- We must be able to restore the thread later
 - Therefore we must save its state before switching to another thread

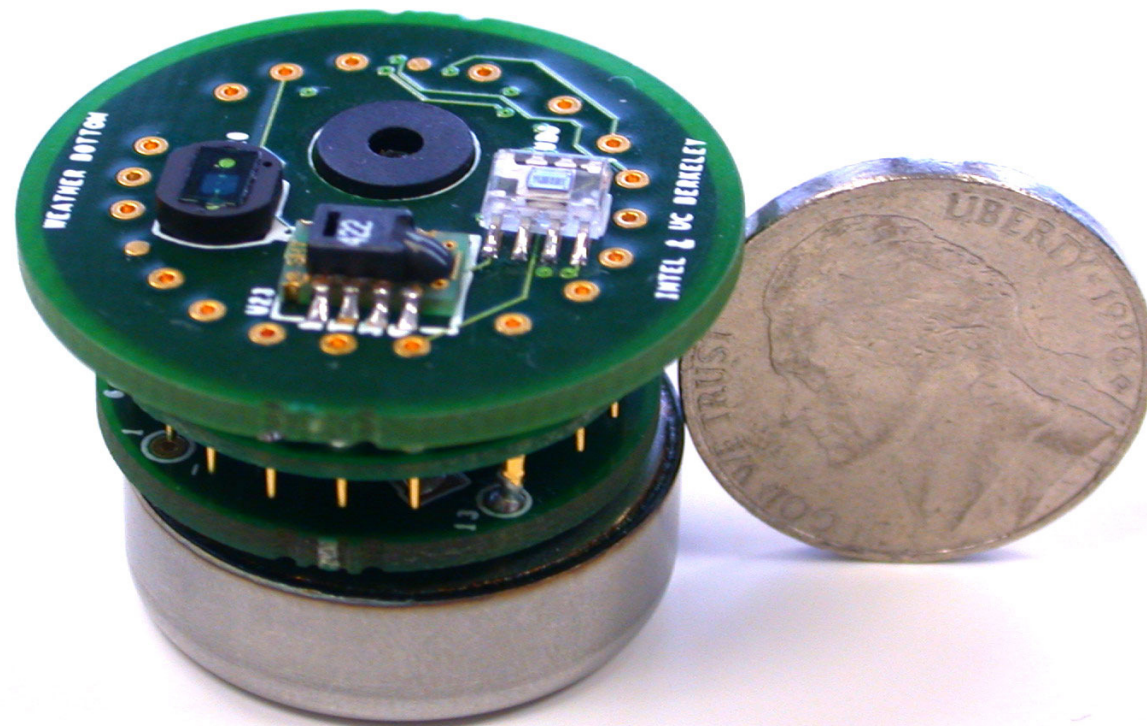
Time sharing systems

- In time sharing systems,
 - Every thread can execute for maximum one round
 - For example, 10msec
 - At the end of the round, the processor is given to another thread



Background on Programming ...

- An Example: Sensor networks ...



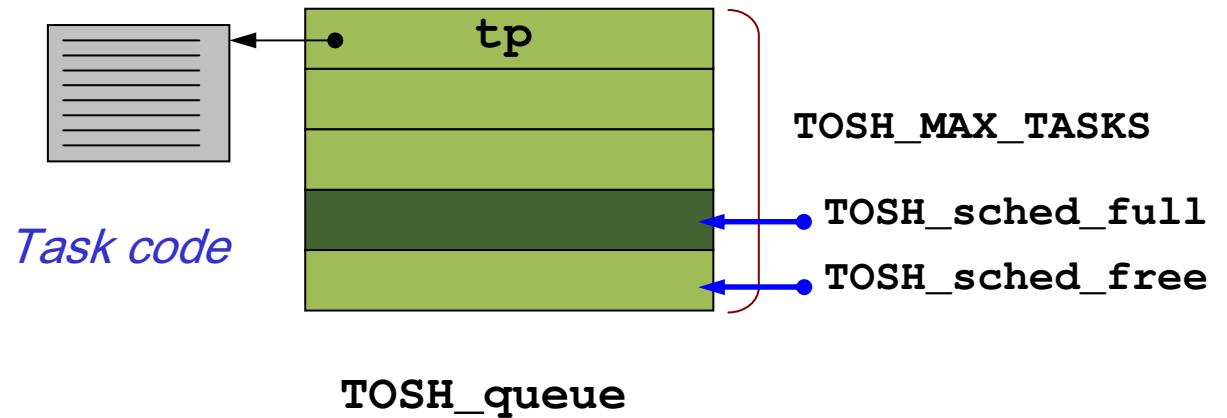
TinyOS: OS for WSN

- **Scheduler:**
 - two level scheduling: events and tasks
 - scheduler is simple FIFO
 - a task can not preempt another task
 - events (interrupts) preempt tasks (higher priority)

```
main {  
    ...  
    while(1) {  
        while(more_tasks)  
            schedule_task;  
        sleep;  
    }  
}
```

TinyOS: OS for WSN

```
typedef struct {  
    void (*tp) ();  
} TOSH_sched_entry_T;
```



```
enum {  
    TOSH_MAX_TASKS = 8,  
    TOSH_TASK_BITMASK = (TOSH_MAX_TASKS - 1)};
```

```
TOSH_sched_entry_T TOSH_queue[TOSH_MAX_TASKS];  
volatile uint8_t TOSH_sched_full;  
volatile uint8_t TOSH_sched_free;
```

TinyOS: OS for WSN

```
void TOSH_sched_init(void)
```

```
{
```

```
    TOSH_sched_free = 0;
```

```
    TOSH_sched_full = 0;}
```

```
bool TOS_empty(void)
```

```
{
```

```
    return TOSH_sched_full == TOSH_sched_free;
```

```
}
```

TinyOS: OS for WSN

```
bool TOS_post(void (*tp) ()) __attribute__((spontaneous)) {
    __nesc_atomic_t fInterruptFlags;
    uint8_t tmp;

    fInterruptFlags = __nesc_atomic_start();

    tmp = TOSH_sched_free;
    TOSH_sched_free++;
    TOSH_sched_free &= TOSH_TASK_BITMASK;

    if (TOSH_sched_free != TOSH_sched_full) {
        __nesc_atomic_end(fInterruptFlags);

        TOSH_queue[tmp].tp = tp;
        return TRUE;
    }
    else {
        TOSH_sched_free = tmp;
        __nesc_atomic_end(fInterruptFlags);

        return FALSE;
    }
}
```

/*
* TOS_post (thread_pointer)
*
* Put the task pointer into the
* next free slot.
* Return 1 if successful,
* 0 if there is no free slot.
*
* This function uses a
* critical section to protect
* TOSH_sched_free.
* As tasks can be posted in both
* interrupt and non-interrupt
* context, this is necessary.
*/

TinyOS: OS for WSN

```
bool TOSH_run_next_task () {
    __nesc_atomic_t fInterruptFlags;  uint8_t old_full;  void (*func)(void);

    if (TOSH_sched_full == TOSH_sched_free) return 0;
    else {
        fInterruptFlags = __nesc_atomic_start();
        old_full = TOSH_sched_full;
        TOSH_sched_full++;
        TOSH_sched_full &= TOSH_TASK_BITMASK;
        func = TOSH_queue[(int)old_full].tp;
        TOSH_queue[(int)old_full].tp = 0;
        __nesc_atomic_end(fInterruptFlags);
        func();
        return 1;
    }
}

void TOSH_run_task() {
    while (TOSH_run_next_task());
    TOSH_sleep();
    TOSH_wait();
}
```

```
/*
 * TOSH_schedule_task()
 *
 * Remove the task at the head of
 * the queue and execute it,
 * freeing the queue entry.
 * Return 1 if a task was executed,
 * 0 if the queue is empty.
 *
 * This function does not need a
 * critical section because it
 * is only run in non-interrupt
 * context; therefore,
 * TOSH_sched_full does not
 * need to be protected.
```


resource

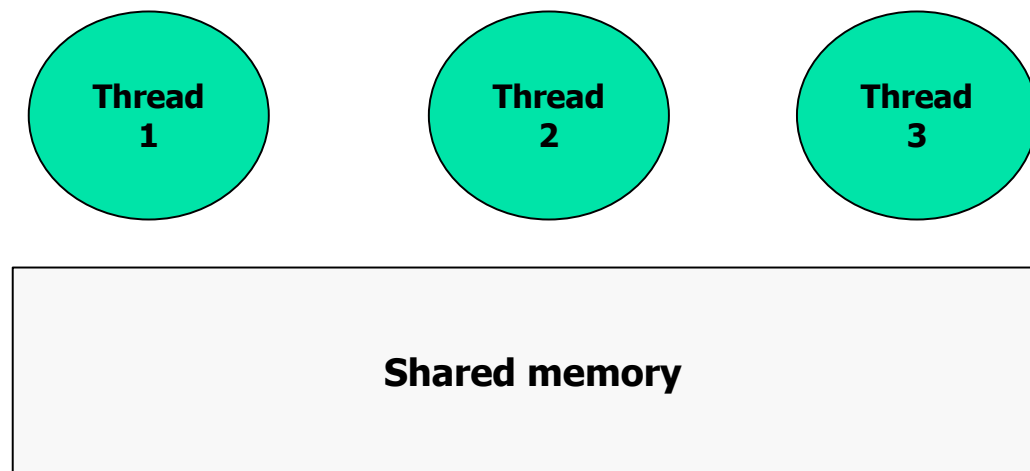
- a resource can be
 - a **HW** resource like a I/O device
 - a **SW** resource, i.e. a data structure
 - in both cases, access to a resource must be regulated to avoid interference
- example 1
 - if two processes want to **print on the same printer**, their access must be sequentialised, otherwise the two printing could be intermingled!
- example 2
 - if two threads **access the same data structure**, the operation on the data must be sequentialized otherwise the data could be inconsistent!

interaction model

- activities can interact according to two fundamental models
 - shared memory
 - All activities access the same memory space
 - message passing
 - All activities communicate each other by sending messages through OS primitives
 - we will analyze both models in the following slides

shared memory

- shared memory communication
 - it was the first one to be supported in old OS
 - it is the simplest one and the **closest to the machine**
 - all threads can access the **same** memory locations



mutual exclusion problem

- we do not know in advance the relative speed of the processes
 - we don't know the order of execution of the hardware instructions

shared memory

```
int x ;
```

```
void *threadA(void *)  
{  
    ...;  
    x = x + 1;  
    ...;  
}
```

```
void *threadB(void *)  
{  
    ...;  
    x = x + 1;  
    ...;  
}
```

- bad interleaving:

...			
LD	R0, x	TA	x = 0
LD	R0, x	TB	x = 0
INC	R0	TB	x = 0
ST	x, R0	TB	x = 1
INC	R0	TA	x = 1
ST	x, R0	TA	x = 1
...			

critical sections

- definitions
 - the **shared object** where the conflict may happen is a “**resource**”
 - the **parts of the code** where the problem may happen are called “**critical sections**”
 - a critical section is a sequence of operations that cannot be interleaved with other operations on the same resource
 - two critical sections on the same resource must be properly sequentialized
 - we say that two critical sections on the same resource must execute in **MUTUAL EXCLUSION**
 - there are three ways to obtain mutual exclusion
 - implementing the critical section as an **atomic operation**
 - **disabling the preemption** (system-wide)
 - **selectively disabling the preemption** (using semaphores and mutual exclusion)

critical sections: atomic operations

- in single processor systems
 - disable interrupts during a critical section
- problems:
 - if the critical section is long, **no interrupt can arrive** during the critical section
 - consider a timer interrupt that arrives every 1 msec.
 - if a critical section lasts for more than 1 msec, a timer interrupt could be lost!
 - **concurrency is disabled** during the critical section!
 - we must avoid conflicts on the resource, not disabling interrupts!

critical sections: atomic operations (2)

- multi-processor
 - define a flag `s` for each resource
 - use `lock(s)/unlock(s)` around the critical section
- problems:
 - **busy waiting**: if the critical section is long, we waste a lot of time
 - cannot be used in single processors!

```
int s;  
...  
lock(s);  
<critical section>  
unlock(s);  
...
```

critical sections: disabling preemption

- single processor systems
 - in some scheduler, it is possible to **disable preemption** for a limited interval of time
 - problems:
 - if a **high priority critical thread needs to execute**, it cannot make preemption and it is delayed
 - even if the high priority task does not access the resource!

<disable preemption>
<critical section>
<enable preemption>

no context
switch may happen
during the critical
section

general mechanism: semaphores


- Dijkstra proposed the **semaphore mechanism**
 - a semaphore is an abstract entity that consists
 - a counter
 - a blocking queue
 - operation wait
 - operation signal
 - the operations on a semaphore are considered atomic

semaphores

- semaphores are basic mechanisms for providing synchronization
 - it has been shown that every kind of synchronization and mutual exclusion can be implemented by using semaphores
 - we will analyze possible implementation of the semaphore mechanism later

```
typedef struct {  
    <blocked queue> blocked;  
    int counter;  
} sem_t;  
  
void sem_init    (sem_t &s, int n);  
  
void sem_wait    (sem_t &s);  
void sem_post    (sem_t &s);
```

Note:
the real prototype
of sem_init is
slightly different!



wait and signal

- a **wait** operation has the following behavior
 - if counter == 0, the requiring thread is blocked
 - it is removed from the ready queue
 - it is inserted in the blocked queue
 - if counter > 0, then counter--;
- a **post** operation has the following behavior
 - if counter == 0 and there is some blocked thread, unblock it
 - the thread is removed from the blocked queue
 - it is inserted in the ready queue
 - otherwise, increment counter

semaphores

```
void sem_init (sem_t *s, int n)
{
    s->count=n;
    ...
}

void sem_wait(sem_t *s)
{
    if (counter == 0)
        <block the thread>
    else
        counter--;
}

void sem_post(sem_t *s)
{
    if (<there are blocked threads>)
        <unblock a thread>
    else
        counter++;
}
```

signal semantics

- what happens when a thread blocks on a semaphore?
 - in general, it is inserted in a BLOCKED queue
- extraction from the blocking queue can follow different semantics:
 - strong semaphore
 - the threads are removed in well-specified order
 - for example, the FIFO order is the fairest policy, priority based ordering, ...
 - signal and suspend
 - after the new thread has been unblocked, a thread switch happens
 - signal and continue
 - after the new thread has been unblocked, the thread that executed the signal continues to execute
- concurrent programs should not rely too much on the semaphore semantic

mutual exclusion with semaphores

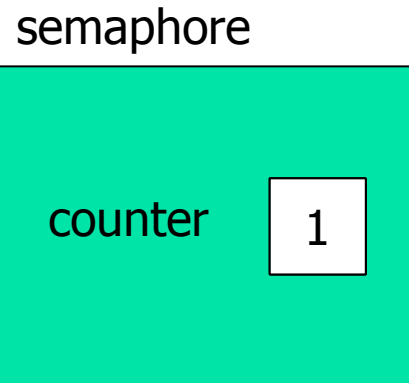
- how to use a semaphore for critical sections
 - define a semaphore **initialized to 1**
 - before entering the critical section, perform a wait
 - after leaving the critical section, perform a post

```
sem_t s;  
...  
sem_init(&s, 1);
```

```
void *threadA(void *arg)  
{  
    ...  
    sem_wait(&s);  
    <critical section>  
    sem_post(&s);  
    ...  
}
```

```
void *threadB(void *arg)  
{  
    ...  
    sem_wait(&s);  
    <critical section>  
    sem_post(&s);  
    ...  
}
```

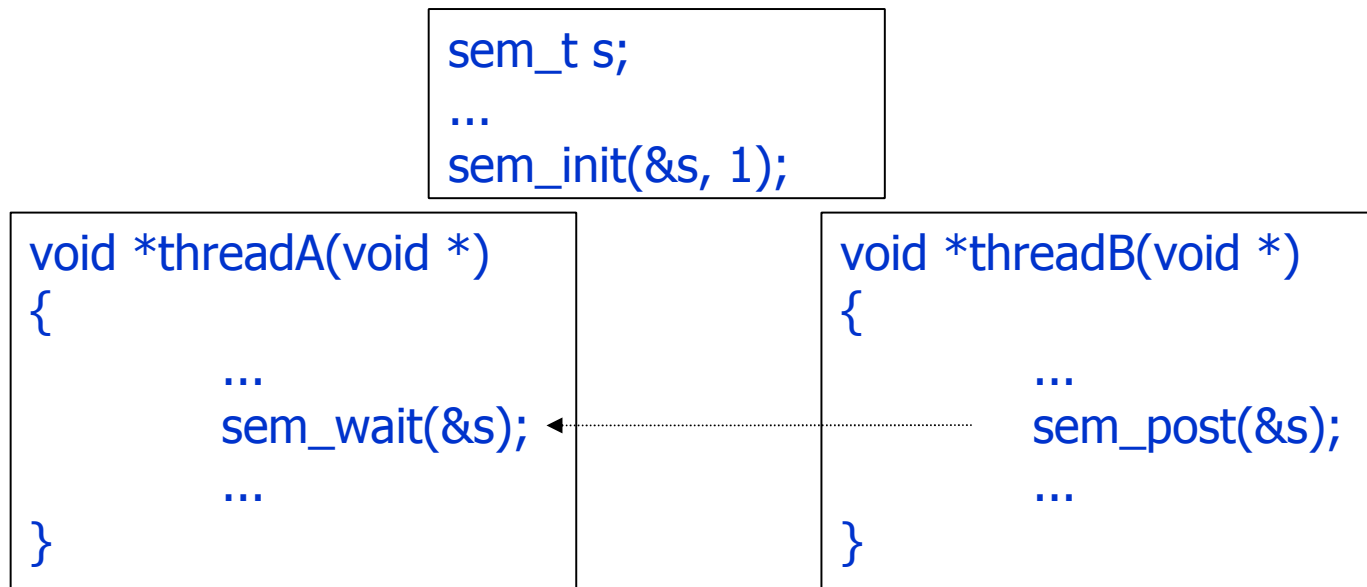
mutual exclusion with semaphores (2)



<code>sem_wait();</code>	(TA)
<code><critical section (1)></code>	(TA)
<code>sem_wait();</code>	(TB)
<code><critical section (2)></code>	(TA)
<code>sem_post();</code>	(TA)
<code><critical section></code>	(TB)
<code>sem_post();</code>	(TB)

synchronization

- how to use a semaphore for synchronization
 - define a semaphore **initialized to 0**
 - at the synchronization point, perform a wait
 - when the synchronization point is reached, perform a post
 - in the example, threadA blocks until threadB wakes it up



- how can both A and B synchronize on the same instructions?

semaphore implementation

- system calls
 - `wait()` and `signal()` involve a possible thread-switch
 - therefore they **must be implemented as system calls!**
 - one blocked thread must be removed from state RUNNING and be moved in the semaphore blocking queue
- protection:
 - a semaphore is itself a shared resource
 - `wait()` and `signal()` are critical sections!
 - they must run with interrupt disabled and by using `lock()` and `unlock()` primitives

semaphore implementation (2)

```
void sem_wait(sem_t *s)
{
    spin_lock_irqsave();
    if (counter==0) {
        <block the thread>
        schedule();
    } else s->counter--;
    spin_lock_irqrestore();
}
```

```
void sem_post(sem_t *s)
{
    spin_lock_irqsave();
    if (counter== 0) {
        <unblock a thread>
        schedule();
    } else s->counter++;
    spin_lock_irqrestore();
}
```

RTOS Standards: POSIX

Industry Insight

Real-Time Linux

Linux 2.6 for Embedded Systems Closing in on Real Time

While not yet ready for hard real-time computing, many new features that make it an excellent platform for embedded computing tasks

by Ravi Gupta, LynuxWorks

With its low cost, abundant features and inherent openness, Linux provides fertile ground for creativity in embedded computing. As its importance grows, we can even expect Linux to become the platform where progress first happens. The question is, could Linux 2.6 be the breakthrough version we've been anticipating for embedded systems—the version that opens the floodgates to Linux acceptance? The answer is “yes.”

The embedded computing universe is vast and encompasses computers of all sizes, from tiny wristwatch cameras to telecommunications switches with thousands of nodes distributed worldwide. Embedded systems can be simple enough to require only small microcontrollers, or they may require massive parallel processors with prodigious amounts of memory and computing power. Linux 2.6 delivers enhancements to provide sup-

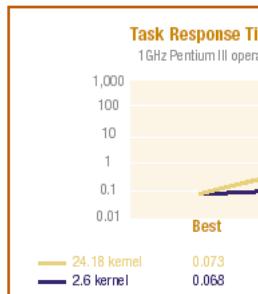


Figure 1 Linux 2.6 shows marked improvement in task response time as loads are increased.

Taken together, these enhancements serve to both “firm up” Linux for embedded computing while making it a more attractive alternative for a wider

The Importance of POSIX

There's a lot to be said about the utility of POSIX. For example, the POSIX standard describes a set of functions for thread creation and management called POSIX threads, or pthreads. This functionality has been available in past versions of Linux, but its implementation has been much improved in 2.6. The Native POSIX Thread Library (NPTL) has been shown to be a significant improvement over the older LinuxThreads approach, and even improves other high-performance alternatives that have been available as patches.

Along with POSIX threads, 2.6 provides POSIX signals and POSIX high-resolution timers as part of the mainstream kernel. POSIX signals are an improvement over UNIX-style signals, which were the default in previous Linux releases. Unlike UNIX signals, POSIX signals cannot be lost and can carry information as an argument. Also, POSIX signals can be sent from one POSIX thread to another, rather than only from process to process like UNIX signals.

Embedded systems also often need to poll hardware or do other tasks on a fixed schedule. POSIX timers make it easy to arrange any task to get scheduled periodically. The clock that the timer uses can be set to tick at a rate as fine as one kilohertz, so that software engineers can control the scheduling of tasks with precision.

This is largely due to its cost-effectiveness in what are often low-margin commodity devices. Linux 2.6 delivers support for several technologies that are key to the success of many types of consumer products. For example, 2.6 includes the Advanced Linux Sound Architecture, or ALSA. This state-of-the-art facility supports USB and MIDI devices with fully thread and multiprocessor-safe software. With ALSA, a system can run multiple sound cards, play and record at the same time or mix multiple audio streams.

USB 2.0 also makes its debut on Linux 2.6. We can expect that high-speed devices will proliferate in the near future, and that Linux will be a leading platform for USB 2.0 products.

These improvements make it a far more worthy platform than in the past.

user interface and sometimes with no operator interface. While previous Linux versions made it possible to build a headless system, some of the support software was not removable, giving the kernel more bulk than was necessary or desirable. Linux 2.6 however can be configured to entirely omit support for unneeded displays, keyboards or mice.

For portable products, Linux 2.6 debuts the Bluetooth wireless interface, which is now taking its place next to 802.11 as a protocol option for wireless communications. With both the SCO datalink for audio and the L2CAP for connection-oriented data transfers available, Linux 2.6 is an excellent choice wherever no fire, short range wireless connectivity

suite.

All Active-enabled tool-

Industry Insight

very large memory sizes have their choice of 64-bit microprocessors with Linux 2.6.

The Intel Itanium 64 architecture was treated in a previous releases of Linux, and support continues in 2.6. Linux 2.6 also continues to cover the AMD64 architecture with support of the AMD Opteron microprocessor. Nor is the PowerPC left out, as PPC64 support is also available. Clearly, as 2.6 illustrates, the Linux community has the momentum to keep up with innovations in large-bus, large-memory computing.

Microcontrollers, on the other hand, have been something of a frontier for Linux. Now they are supported on the mainstream Linux 2.6 kernel. In most cases, previous instances of Linux required a full-featured microprocessor with a memory management unit (MMU). But simpler microcontrollers are typically the more appropriate choice when low cost and simplicity are called for.

There have been ways to put Linux on MMU-less processors prior to version 2.6. The Linux for Microcontrollers project has been a successful branch of Linux for some important small systems. Version 2.6 integrates a significant portion of uClinux into the production kernel, bringing microcontroller support into the Linux mainstream.

The Linux 2.6 version supports several current microcontrollers that don't have memory management units. These include Motorola m68k processors such as Dragonball and ColdFire, as well as Hitachi H8/300 and NEC v850 microprocessors. The ETRAX family of networking microcontrollers by Axis Communications is also supported.

As a caveat, Linux running on MMU-less processors will still be multitasking, but will obviously not have the memory protection provided on fully endowed processors. Consistent with the lack of true processes on these small platforms, there is also little in the way of security.

RTLinux was one of the

RTOS Standards: OSEK

The image shows a presentation slide titled "RTA Software Products Overview" overlaid on a browser window displaying the Metrowerks website. The slide content is as follows:

RTA Software Products

Overview

The RTA product family is made up of tools and software components for developing optimized embedded real-time systems. As the required functionality for ECUs becomes ever more demanding, the RTA product family offers the ideal solution to deliver complex real-time software systems on time and to budget.

Closed-Loop Development

The diagram shows a cycle of three tools: RTA-OSEK Planner, RTA-OSEK Builder, and RTA-OSEK Competent. Arrows indicate a clockwise flow: Planner to Builder, Builder to Competent, and Competent back to Planner. A dashed arrow also points from Planner to Builder.

Use of the OSEK operating system (OS) standard is accepted practice across the worldwide automotive industry. RTA-OSEK Component is the world's best implementation of the OSEK OS standard that has been refined and enhanced as a result of many years experience with successful ECU projects. It offers full compliance with OSEK OS features and supports a wide range of microcontrollers that are commonly used for automotive applications.

A key benefit of RTA-OSEK is the ability to use its Planner tool to model an application's real-time performance and analyze whether all of the associated real-time performance requirements will be met. In this way, the application code can be written with the confidence that costly reworking to avoid performance problems will not be necessary.

The Builder tool of RTA-OSEK permits the configuration of every aspect of the OSEK OS application. Using this information, the Builder is able to produce highly optimized OSEK OS implementation specifically for the configured application.

RTA-OSEK Planner and Builder integrate tightly into the software design process, with both graphical and command line modes of operation.

When a working system is available, the powerful features of

Definitions of Real-time system

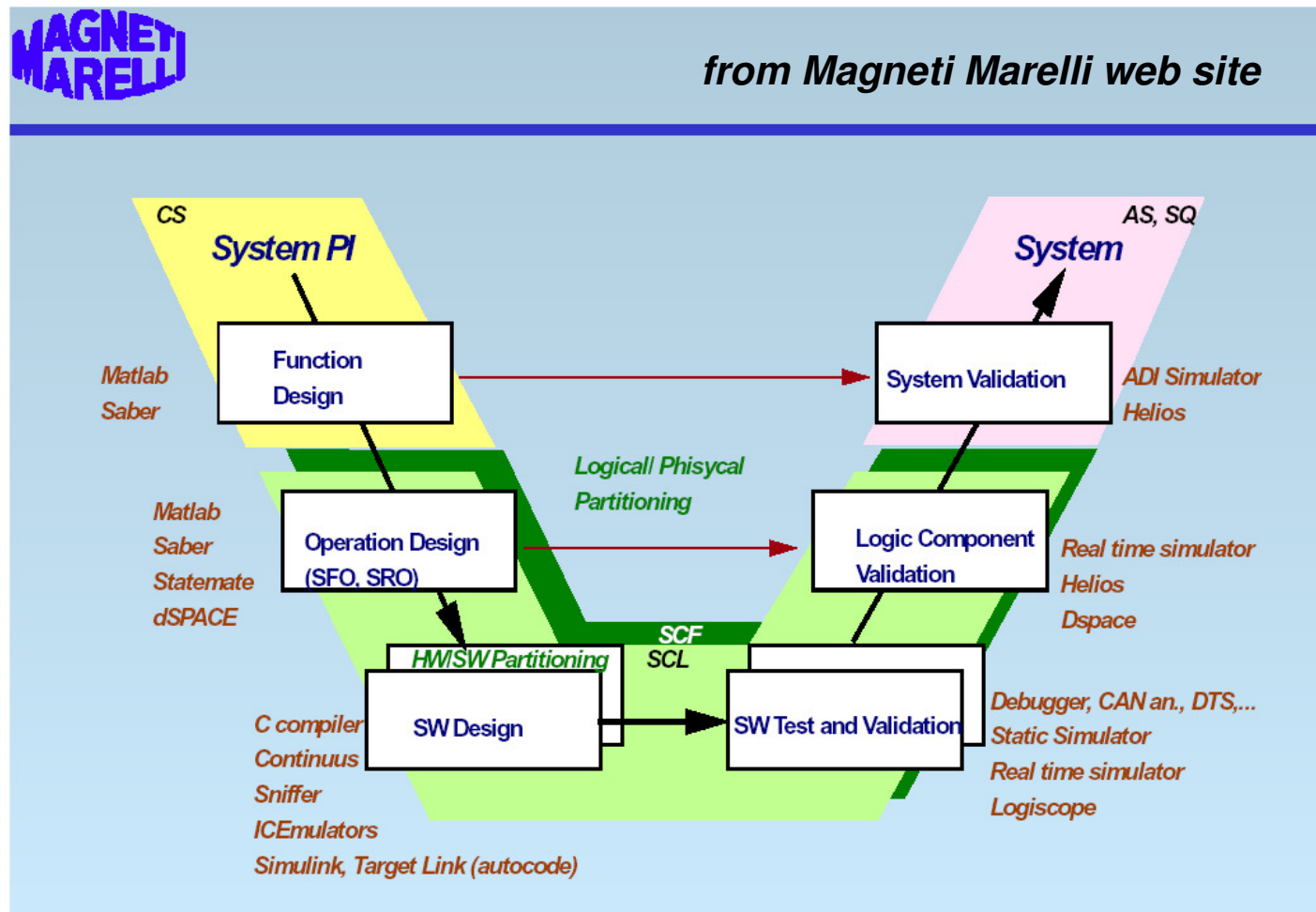
- *Interactions between the system and the environment (environment dynamics).*
- *Time instant when the system produces its results (performs an action).*
- A real-time operating system is an interactive system that maintains an ongoing relationship with an asynchronous environment i.e. an environment that progresses irrespective of the RTS
- A real-time system responds in a (timely) predictable way to (un)predictable external stimuli arrival.
- (Open, Modular, Architecture Control user group - OMAC): a hard real-time system is a system that would fail if its timing requirements were not met; a soft real-time system can tolerate significant variations in the delivery of operating system services like interrupts, timers, and scheduling.
- **In real-time computing correctness depends not only on the correctness of the logical result of the computation but also on the result delivery time (timing constraints).**

Timing constraints

- Where do they come from ?
- From system specifications (design choices?)

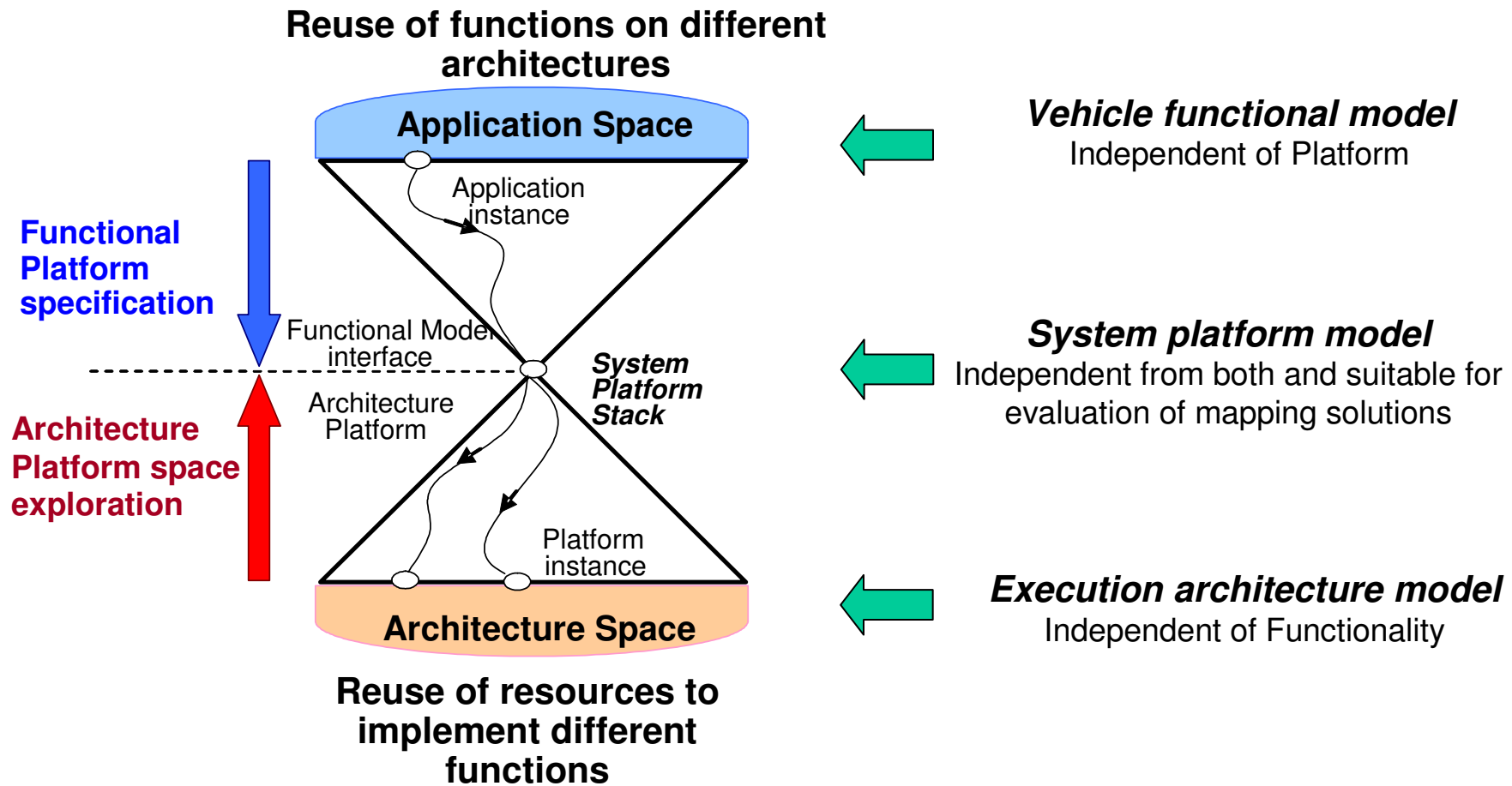
Automotive systems development process

- V-cycle



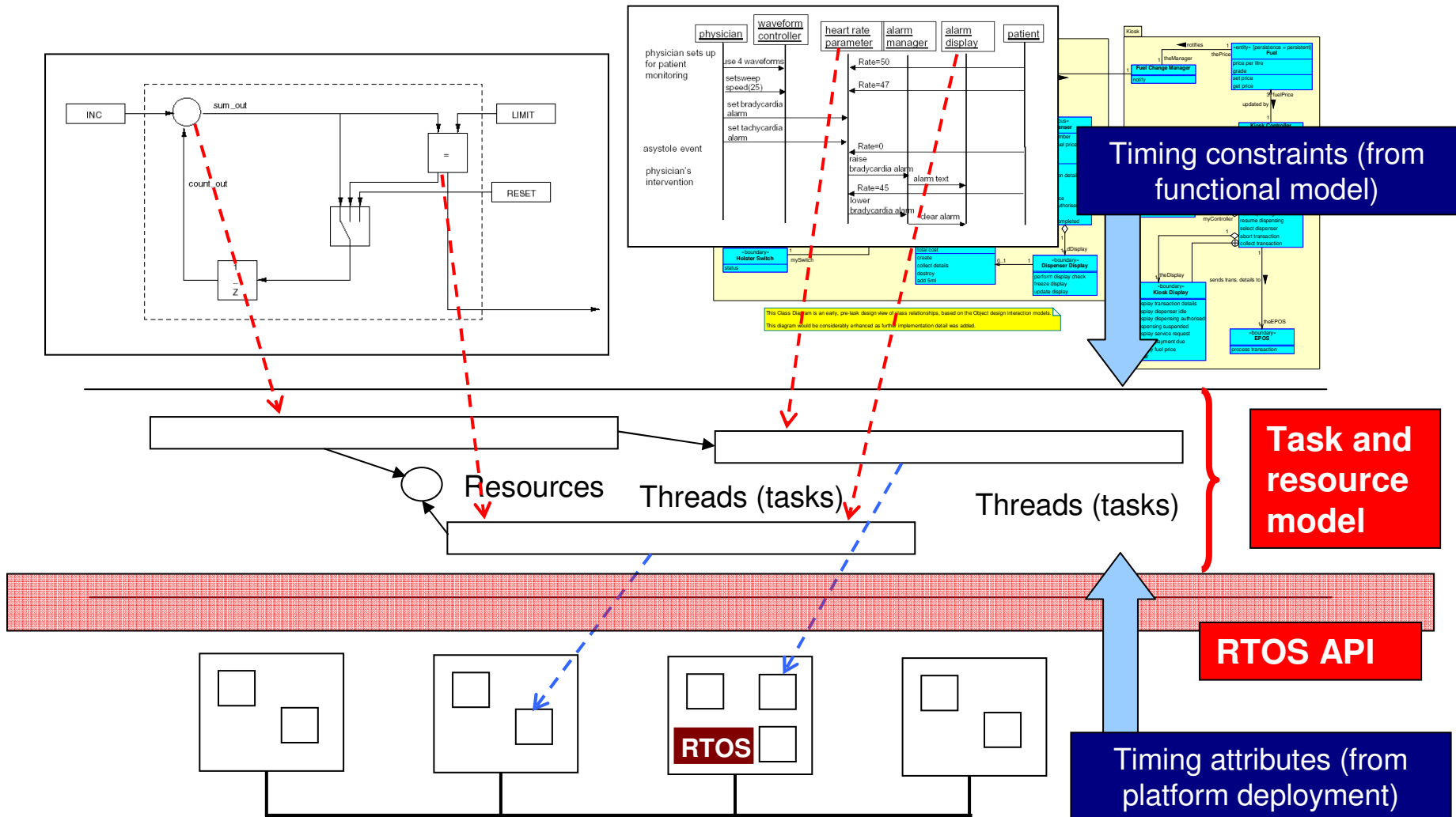
RTS and Platform-Based Design

Platform based design and the decoupling of **Functionality** and **Architecture** enable the reuse of components on both ends in the meet-in-the middle approach



RTS and Platform-Based Design

- Design (continued): matching the logical design into the SW architecture design



An introduction to Real-Time scheduling

- Application of schedulability theory (worst case timing analysis and scheduling algorithms)
- for the development of scheduling and resource management algorithms inside the RTOS, driving the development of efficient (in the worst case) and predictable OS mechanisms (and methods for accessing OS data structures)
- for the evaluation and later verification of the design of embedded systems with timing constraints and possibly for the synthesis of an efficient implementation of an embedded system (with timing constraints) model
 - synthesis of the RTOS

Real-time scheduling

- Assignment of system resources to the software threads
- System resources
 - physical: CPU, network, I/O channels
 - logical: shared memory, shared mailboxes, logical channels
- Typical operating system problem
- In order to study real-time scheduling policies we need a model for representing
 - abstract entities
 - actions,
 - events,
 - time, timing attributes and constraints
 - design entities
 - units of computation
 - mode of computation
 - resources

Classification of RT Systems

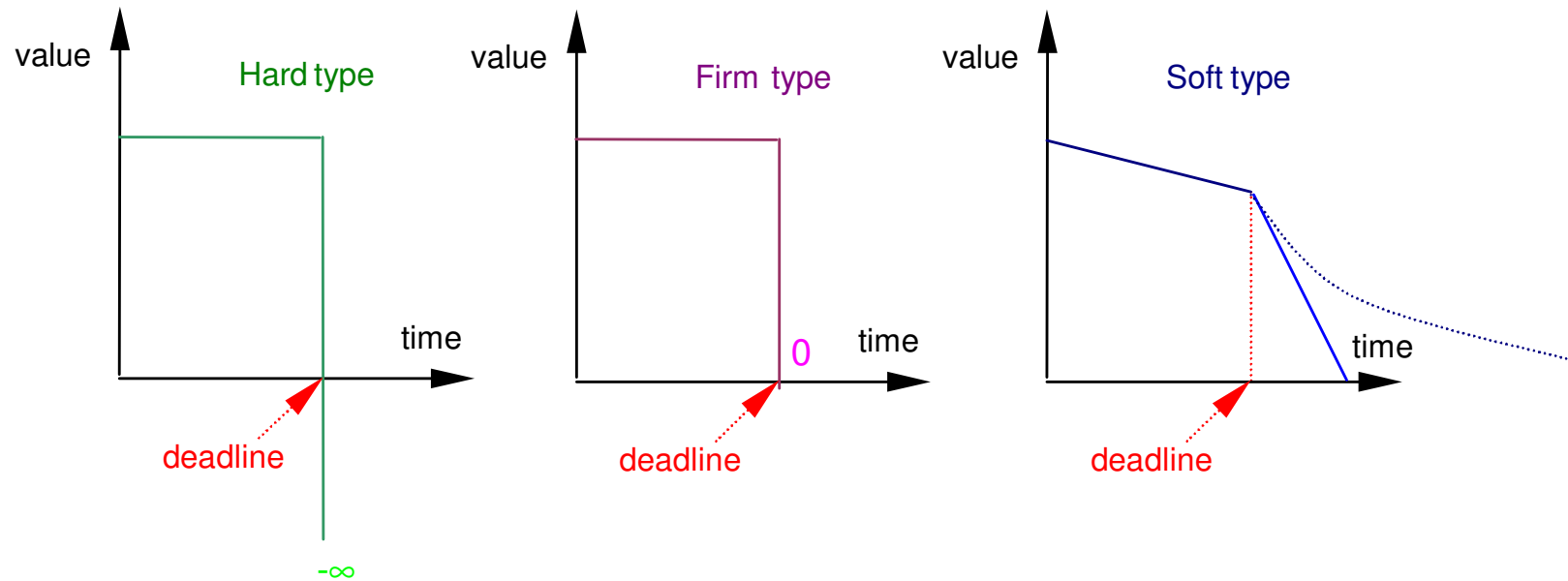
- Based on input
 - *time-driven*: continuous (synchronous) input
 - *event-driven*: discontinuous (asynchronous) input
- Based on criticality of timing constraints
 - *hard RT systems*: response of the system within the timing constraints is crucial for correct behavior
 - *soft RT systems*: response of the system within the timing constraints increases the value of the system
- Based on the nature of the RT load:
 - *static*: predefined, constant and deterministic load
 - *dynamic*: variable (non deterministic) load
- real world systems exhibit a combination of these characteristics

Classification of RT Systems: criticality

- Typical Hard real time systems
 - Aircraft, Automotive
 - Airport landing services
 - Nuclear Power Stations
 - Chemical Plants
 - Life support systems
- Typical Soft real time systems
 - Multimedia
 - Interactive video games

Classification of RT Systems: criticality

- Hard, Soft and Firm type

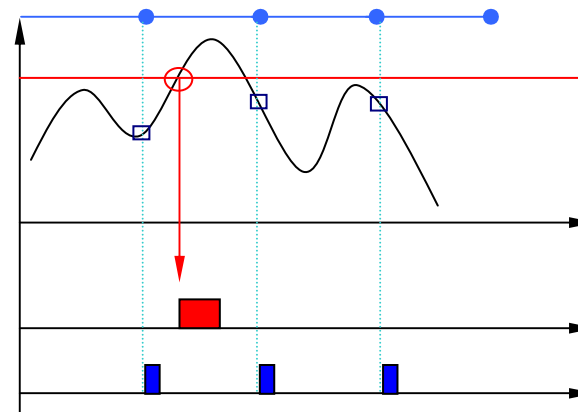
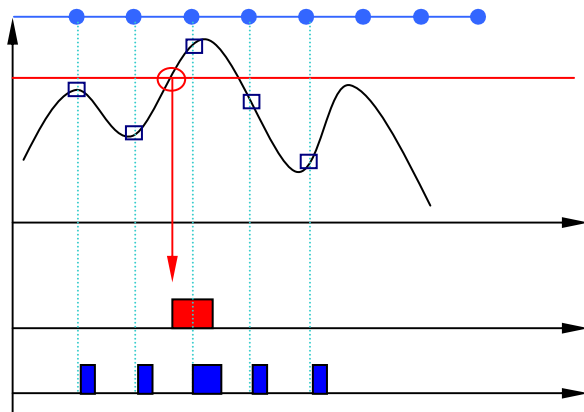


Classification of RT Systems: Input-based

- Event-Triggered vs. Time-Triggered models
- Time triggered
 - Strictly periodic activities (periodic events)
- Event triggered
 - activities are triggered by external or internal asynchronous events, not necessarily related to a periodic time reference

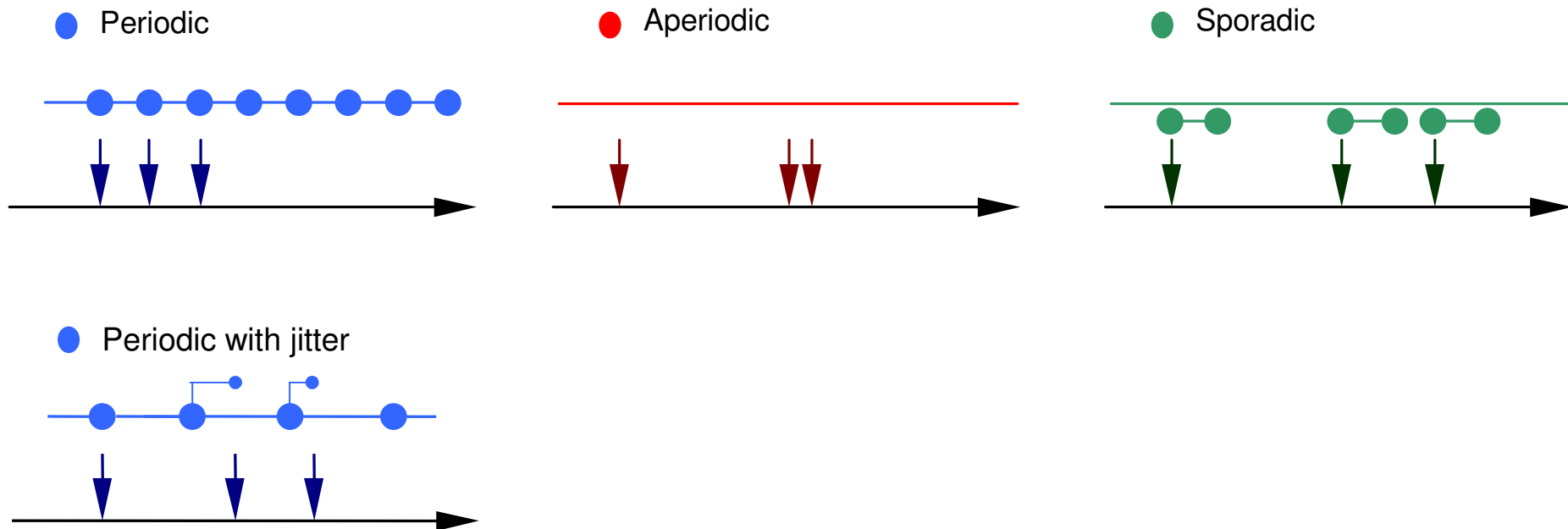
Classification of RT Systems: Input-based

- Example, activity to be executed when the temperature exceeds the *warn* level:
- *event triggered*
 - Action triggered only when temperature $>$ *warn*
- *time triggered*
 - controls temperature every *int* time units; recovery is triggered when temperature $>$ *warn*



Classification of RT Systems: Input-based

- Activation models

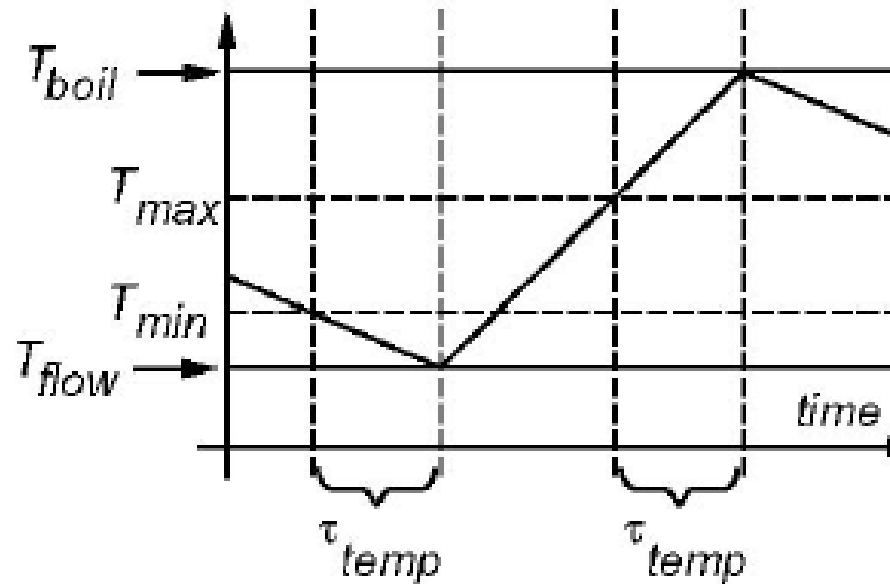


Modeling Real-time systems

- We need to identify (in the specification and design phase)
 - Events and Actions (System responses).
- Some temporal constraints are explicitly expressed as a results of system analysis
 - “The alarm must be delivered within 2s from the time instant a dangerous situation is detected”
- More often, timing constraints are hidden behind sentences that are apparently not related to time ...
 - And are the result of design choices

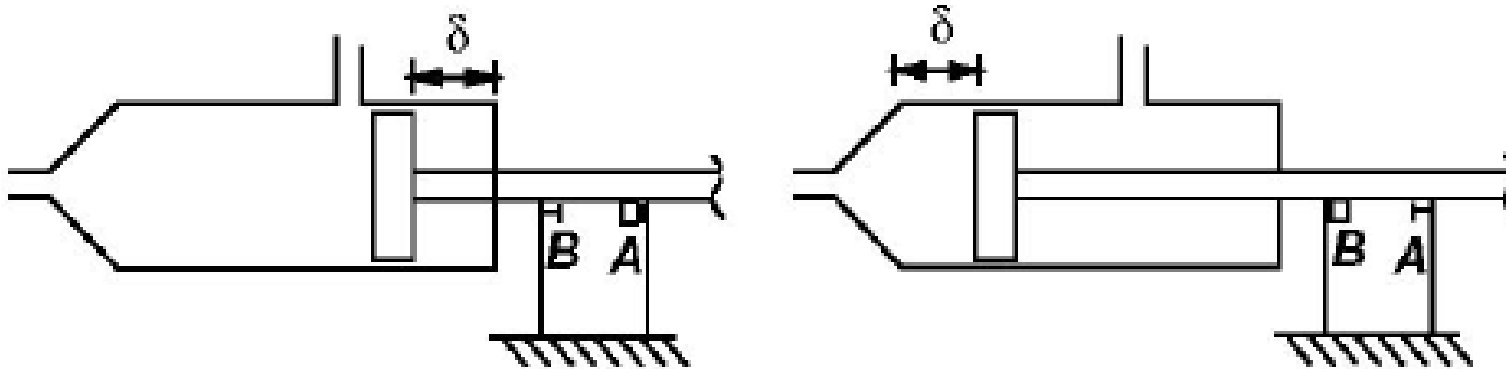
Modeling Real-time systems

- Example: plastic molding
- The controller of the temperature must be activated within τ_{temp} seconds from the time instant when temperature thresholds are surpassed, such that $T_{boil} < T < T_{flow}$



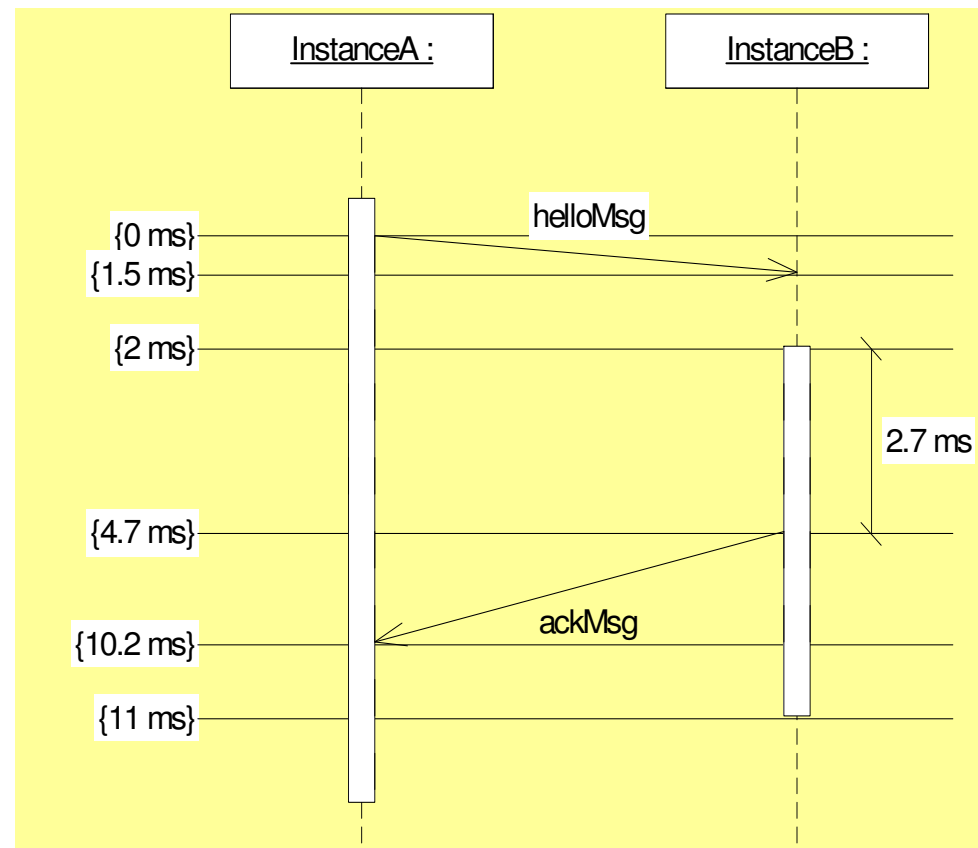
Modeling Real-time systems

- ... the injector must be shut down no more than τ_{inj} seconds after receiving the end-run signals A or B such that $v_{inj}\tau_{inj} < \delta$



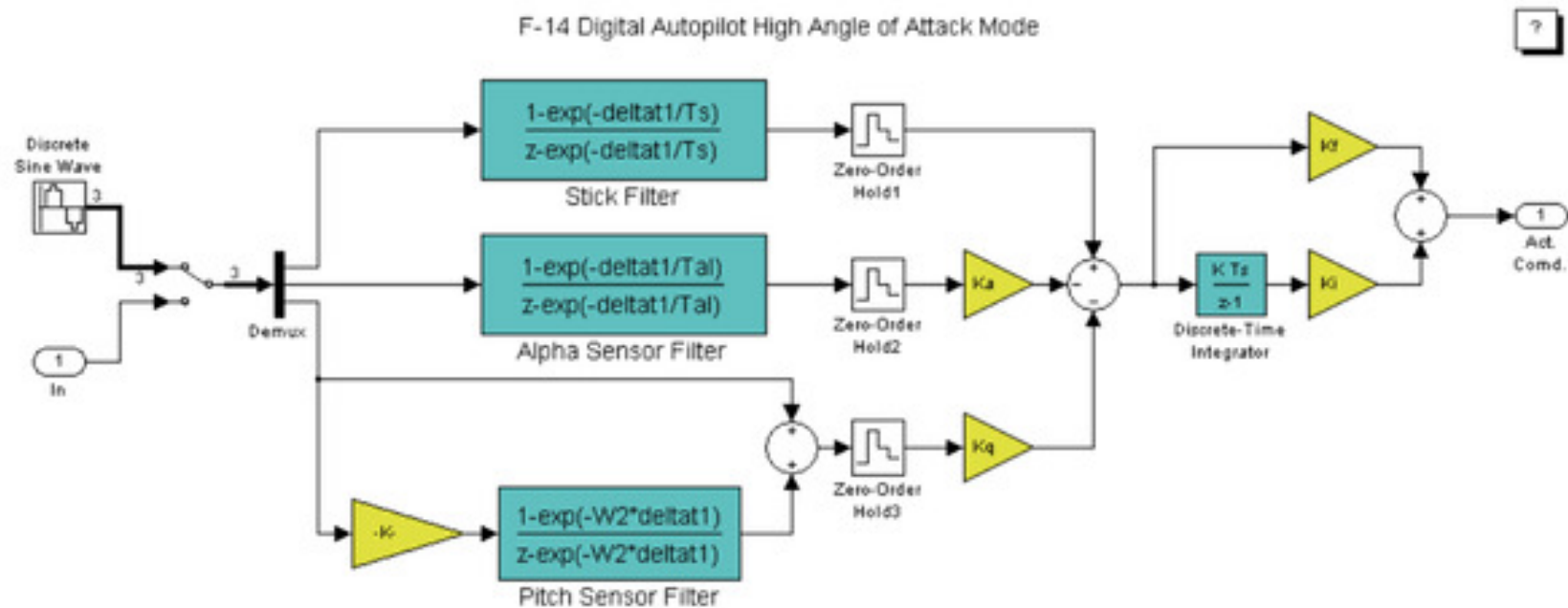
Modeling Real-time systems

- (UML profile, alternate notation)



Modeling Real-time systems

- What type of timing constraints are in a Simulink diagram?

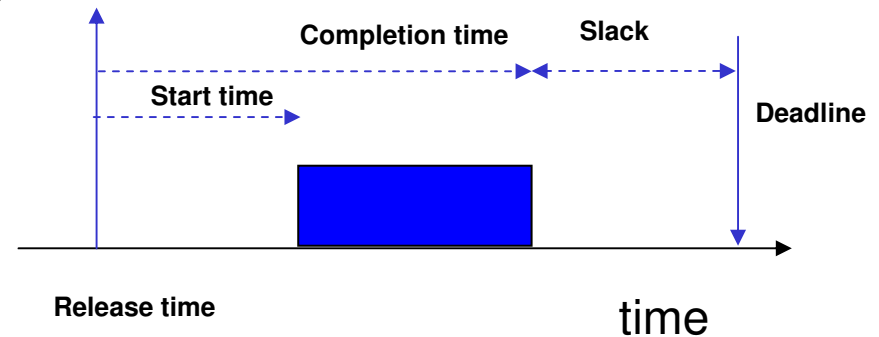


Scheduling of Real-time systems

- What are the key concepts for real-time systems?
 - Schedulable entities (threads)
 - Shared resources (physical – HW / logical)
 - Resource handlers (RTOS)
- Defined in the design of the *Architectural level*

Our definition of real-time

- Based on timing correctness
 - includes timing constraints
 - Response times
 - Deadlines
 - Jitter
 - Release times, slack ...
- Precedence and resource constraints

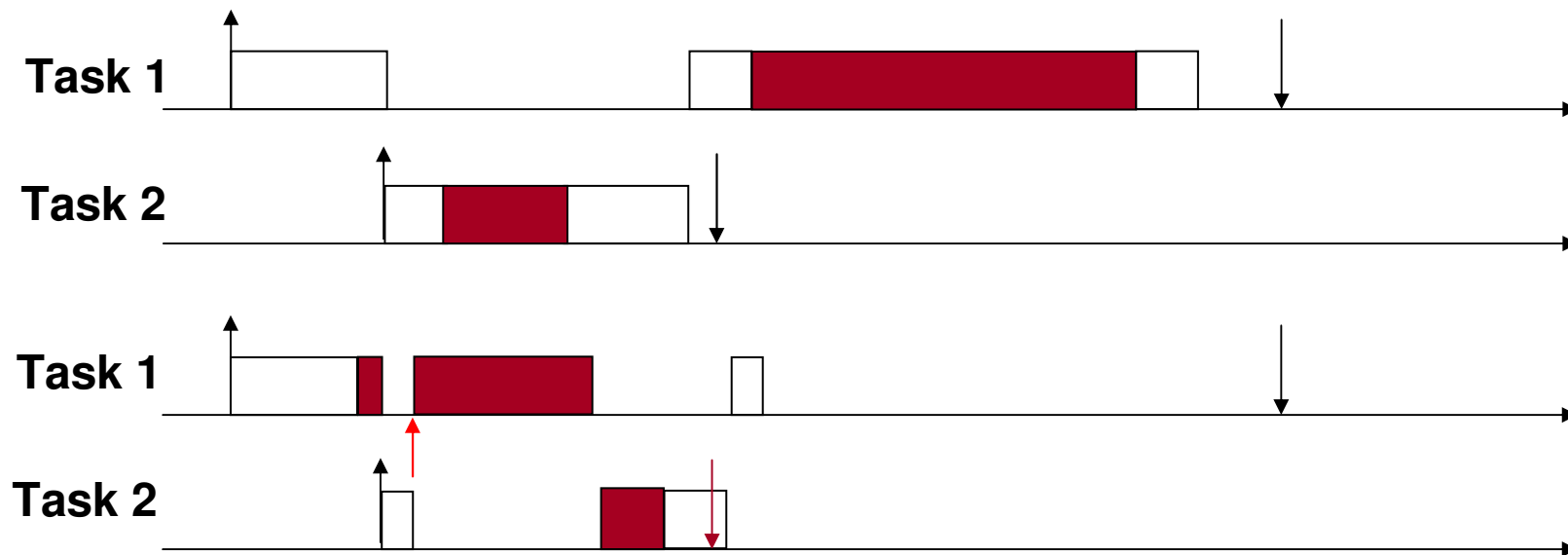


Real-time systems: handling timing constraints

- Real Time = the fastest possible implementation dictated by technology and/or budget constraints ?
- “the fastest possible response is desired. But, like the cruise control algorithm, fastest is not necessarily best, because it is also desirable to keep the cost of parts down by using small microcontrollers. What is important is for the application requirements to specify a worst-case response time. The hardware and software is then designed to meet those specifications“
- “Embedded systems are usually constructed with the least powerful computer that can meet the performance requirements. Marketing and sale concerns push for using smaller processors and less memory reducing the so-called recurring costs”

Real-time systems: handling timing constraints

- Faster is always better ?

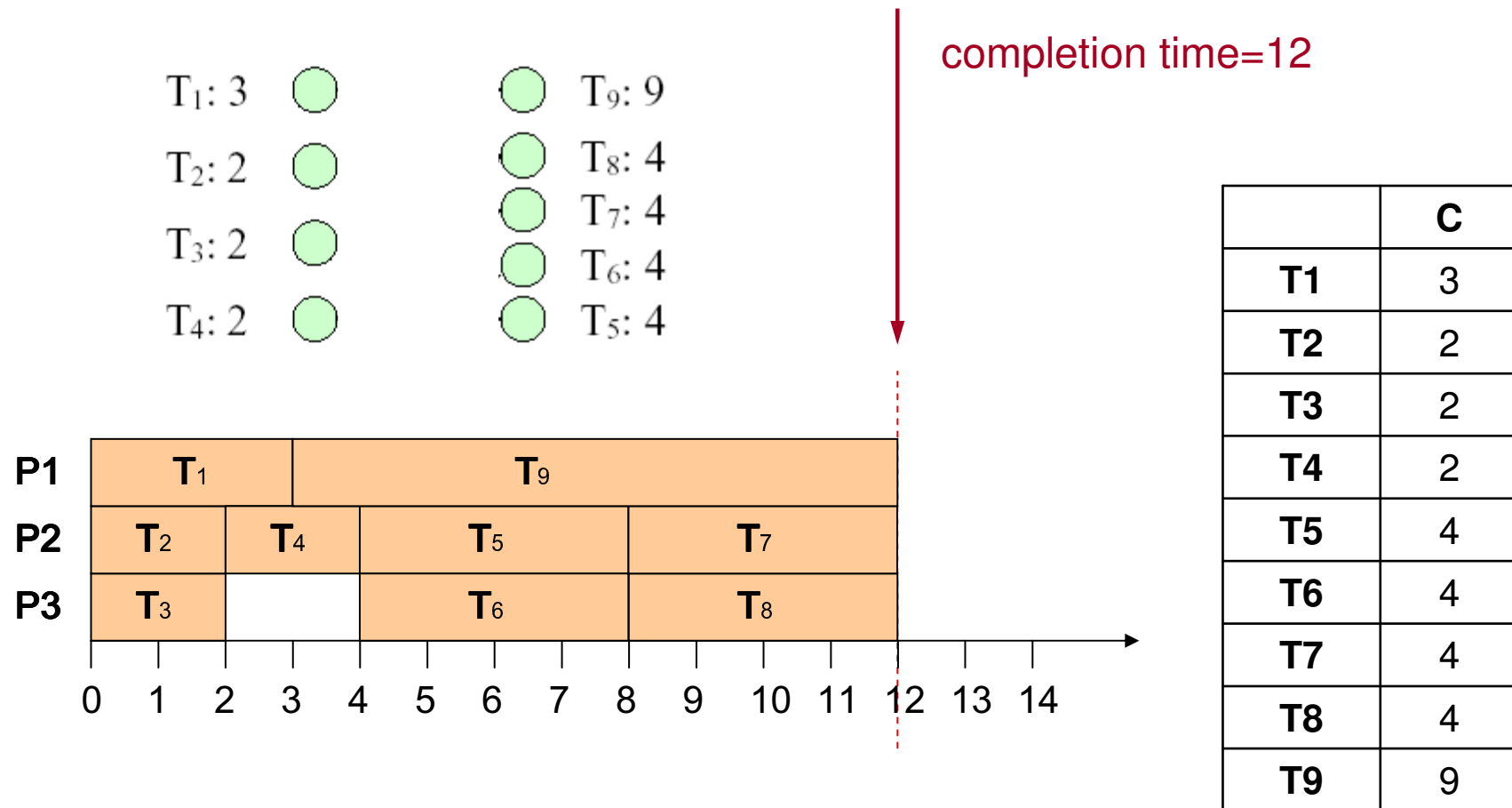


Real-time systems: handling timing constraints

- Scheduling anomalies [Richards]
- The response time of a task in a real-time system may **increase** if
 - the execution time of some of the tasks is reduced
 - precedence constraints are removed from the specifications
 - additional resources (processors) are added to the system ...

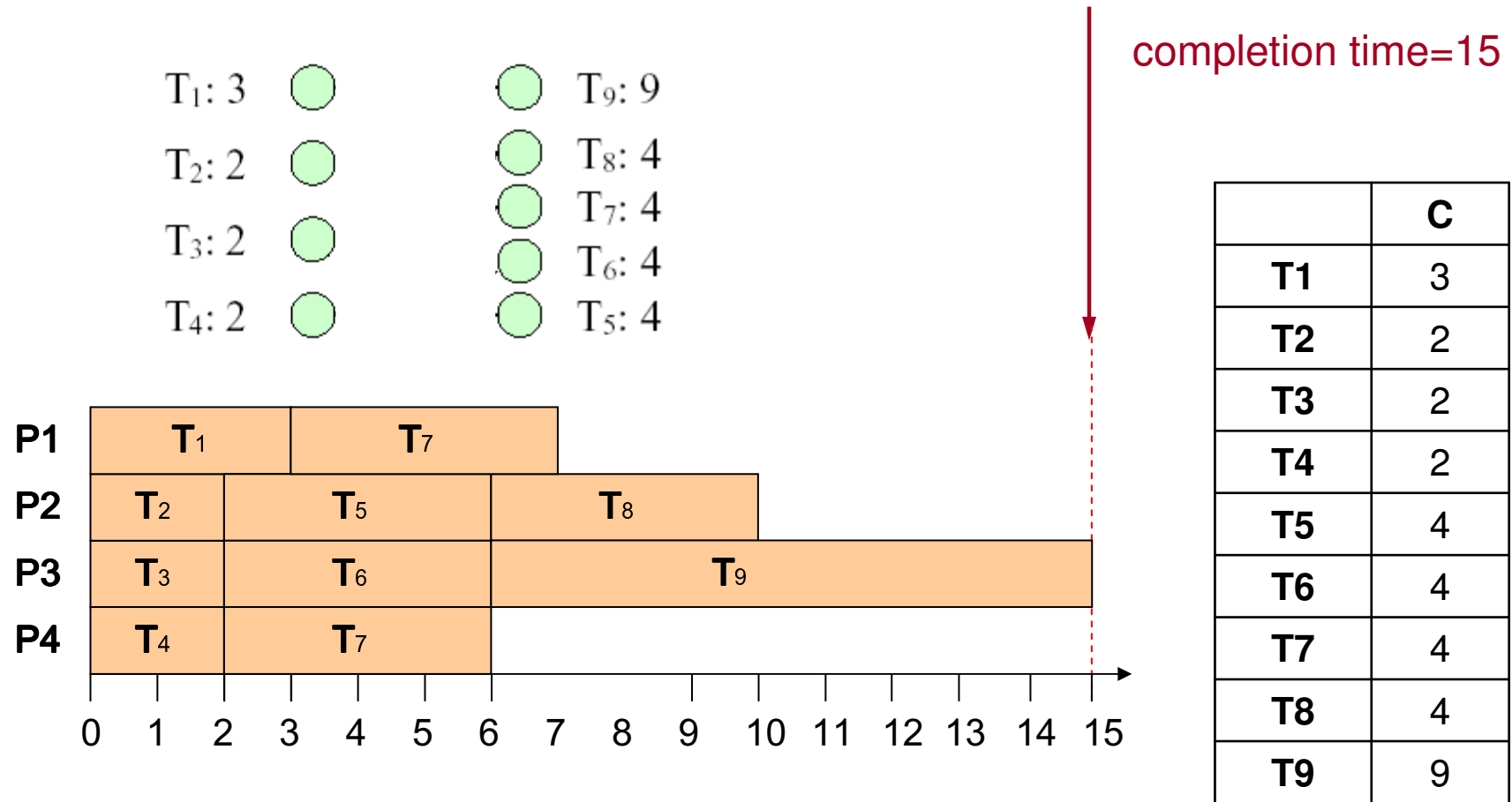
Operating Systems background

- Scheduling anomalies



Operating Systems background

- Increasing the number of processors ...

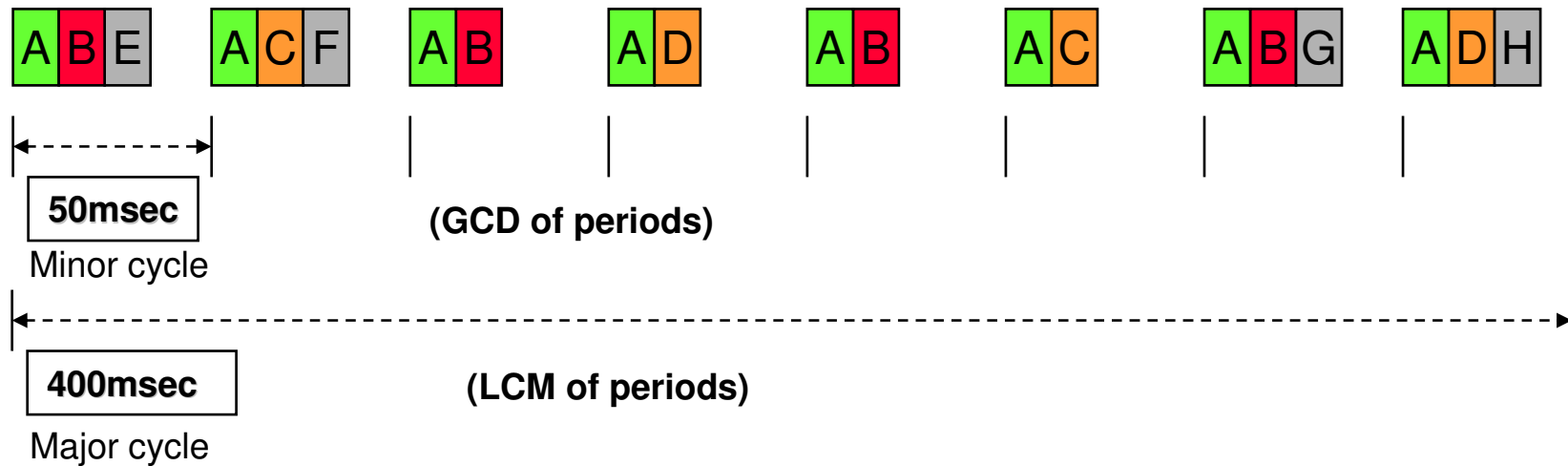






A typical solution: cyclic scheduler

- Used for now 40 years in industrial practice
 - military
 - navigation
 - monitoring
 - control ...
- Examples
 - space shuttle
 - Boeing 777
 - code generated by Mathworks embedded coder (single task mode)

A typical solution: cyclic scheduler

The individual tasks/functions are arranged in a cyclic pattern according to their rates, the schedule is organized in a major cycle and a minor cycle.



-  = 50 msec function A
-  = period 100 msec (function B)
-  = 200 msec (2 functions C, D)
-  = 400 msec (4 functions : E, F, G, H)

A typical solution: cyclic scheduler

Advantages:

- simplicity (no true OS, only dispatcher tables)
- efficiency
- observability
- jitter control
- extremely general form (handles general precedence and resource constraints)

Disadvantages

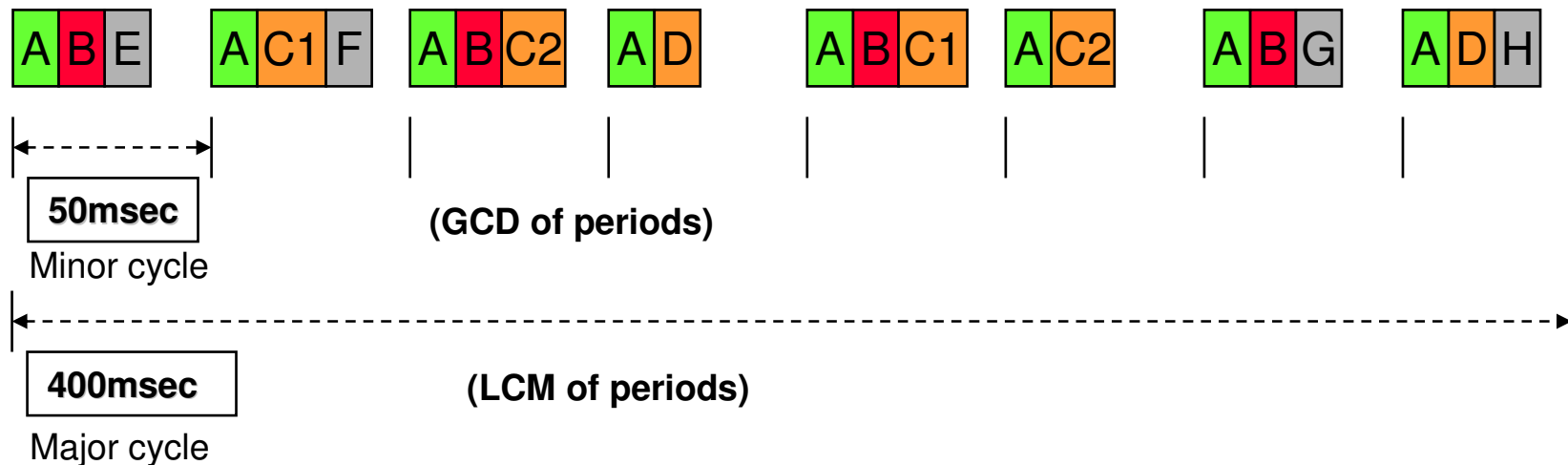
- almost no flexibility
- potentially hides fundamental information on task interactions
- additional constraints on scheduling
 - all functions scheduled in a minor cycle must terminate before the end of the minor cycle.
 $A+B+E \leq \text{minor cycle time}$ (the same for $A+C+F$, $A+B+G$, $A+D+H$)

Problems with cyclic schedulers

- Solution is customized upon the specific task set
 - a different set, even if obtained incrementally, requires a completely different solution
- Race conditions may be “hidden” by the scheduling solution
 - see the shared resource section
 - problems due to non-protected concurrent accesses to shared resources may suddenly show up in a new solution

Problems with cyclic schedulers

- Solution is customized upon the specific task set
- what happens if in our example ...
 - we change the implementation of C and $A+C+F > \text{minor cycle}$?
 - Possible solution: C is split in C1 and C2 (this might not be easy)



- we change the execution rate of (some) functions?
- The minor and major cycle time change! We must redo everything!

From cyclic schedulers (Time triggered systems) to Priority-based scheduling

- Periodic timer:
 - once initialized send periodic TimeEvents at the appropriate time instants (minor cycle time) until explicitly stopped or deleted
- Threads exclusively activated by periodic timers are periodic tasks
 - scheduled according to a fixed priority policy

A Taxonomy for FP scheduling

	D=T	D≤T	Any D
Optimal pri ass.	Y (RM) [L&L]	Y (DM) [JP] [Leh]	Yes (Aud) [Aud]
Test	Util (sufficient) Resp. time (1st inst.) Process. demand	Resp. time (1st inst.) Process. demand	Resp. time (1st busy period) Process. demand
With Resources			
Optimal Pri ass.	NP-complete [Mok]		
Test	Sufficient tests (PIP,PCP)		
With Offsets			
Optimal Pri ass.	Yes (Aud) [Aud]		
Test	Processor Demand		

Case 1: Independent periodic tasks

- Activation events are **periodic (period=T)**,
- **Deadlines** are timing constraints on the execution of tasks (**D=T**)
 - every task instance must be completed before the next instance
 - (no need to provide queues (buffers) for activation events)
- tasks are **independent**
 - the execution of a task does not depend upon the execution (completion) of another task
 - periods may be correlated
- The **execution time** of each task is constant
 - approximated with the worst case execution time

Task set

- n independent tasks $\tau_1, \tau_2, \dots, \tau_n$
- Task periods T_1, T_2, \dots, T_n
 - the activation rate of τ_i is $1/T_i$
- Execution times are C_1, C_2, \dots, C_n

Scheduling algorithm

- Rules dictating the task that needs to be executed on the CPU at each time instant
- preemptive & priority driven
 - task have priorities
 - statically (design time) assigned
 - at each time the highest priority task is executed
 - if a higher priority task becomes ready, the execution of the running task is interrupted and the CPU given to the new task
- In this case, **scheduling algorithm = priority assignment** + priority queue management

Priority-based scheduling

- Static (fixed priorities)
- as opposed to ... Dynamic
 - the priority of each task instance may be different from the priority of other instances (of the same task)

Definitions ...

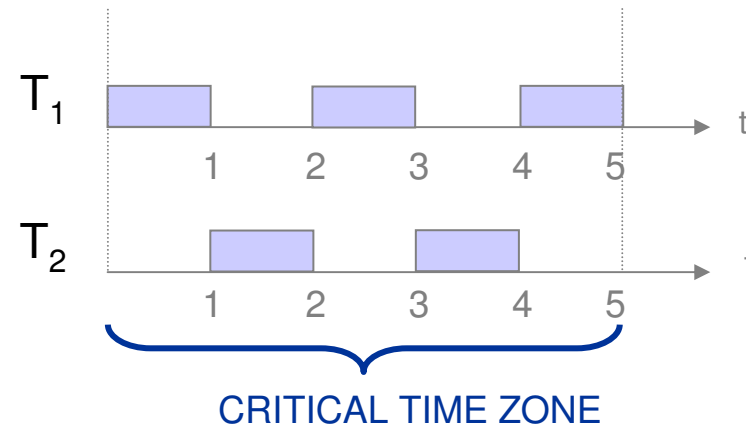
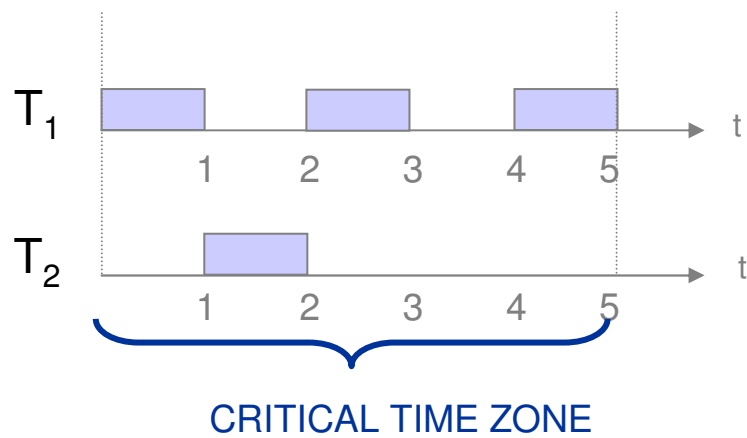
- **Deadline** of a task - latest possible completion time and time instant of the next activation
- **Time Overflow** when a task completes after the deadline
- A scheduling algorithm is **feasible** if tasks can be scheduled without overflow
- **Critical instant** of a task = time instant t_0 such that, if the task instance is released in t_0 , it has the worst possible response (completion) time (Critical instant of the system)
- **Critical time zone** time interval between the critical instant and the response (completion) of the task instance

Critical instant for fixed priorities

- *Theorem 1: the critical instant for each task is when the task instance is released together with (at the same time) all the other higher priority instances*
- The critical instant may be used to check if a priority assignment results in a feasible scheduling
 - if all requests at the critical instant complete before their deadlines

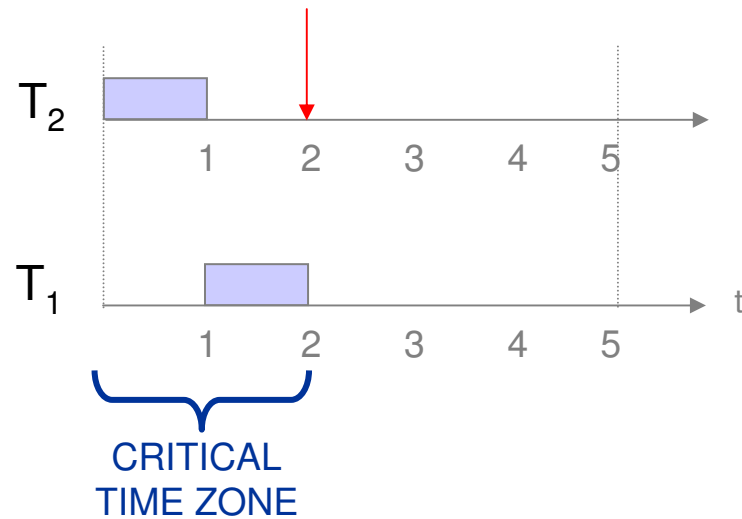
Example

- τ_1 & τ_2 with $T_1=2$, $T_2=5$, & $C_1=1$, $C_2=1$
- τ_1 has higher priority than τ_2
 - priority assignment is feasible
 - C_2 may be increased to 2 and the task set is still feasible



Example

- However, if τ_2 has higher priority than τ_1
 - Assignment is still feasible
 - but computation times cannot be further increased $C_1=1, C_2=1$



Rate Monotonic

- Priority assignment rule **Rate-Monotonic (RM)**
- Assign priorities according to the activation rates (independently from computation times)
 - higher priority for higher rate tasks (hence the name rate monotonic)
- RM is optimal (among all possible static priority assignments)
- *Theorem 2: if the RM algorithm does not produce a feasible schedule, then there is no fixed priority assignment that can possibly produce a feasible schedule*

A priori guarantees

- Understanding at design time if the system is schedulable
- different methods
 - utilization based
 - based on completion time
 - based on processor demand

Processor Utilization

- Processor Utilization Factor: fraction of processor time spent in executing the task set
 - i.e. 1 - fraction of time processor is idle
- For n tasks, $\tau_1, \tau_2, \dots, \tau_n$ the utilization factor U is

$$U = C_1/T_1 + C_2/T_2 + \dots + C_n/T_n$$

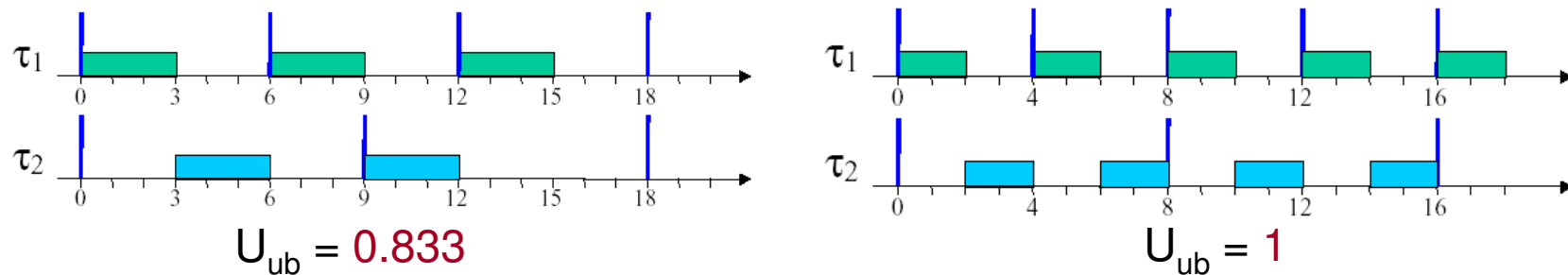
- U can be improved by increasing C_i 's or decreasing T_i 's as long as tasks continue to satisfy their deadlines at their critical instants

Processor Utilization

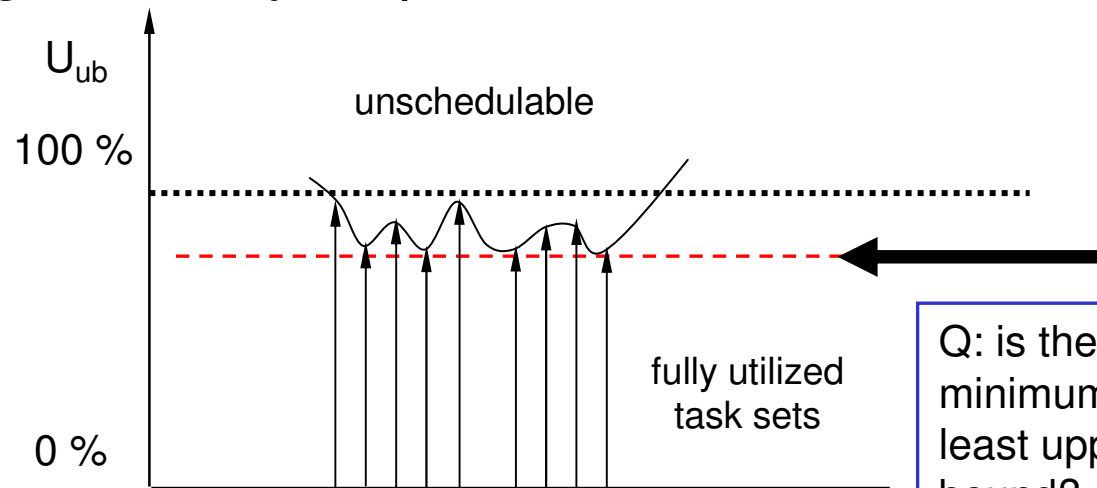
- Given a priority assignment, a set of tasks **fully utilizes** a processor if:
 - the priority assignment is feasible for the set
 - and, if an increase in the run time of any task in the set will make the priority assignment infeasible
- The **least upper bound of U** is the minimum of the U's over all task sets that fully utilize the processor
 - for all task sets whose U is below this bound, \exists a fixed priority assignment which is feasible
 - U above this bound can be achieved only if the task periods T_i 's are suitably related

Processor Utilization

- The upper bound on U depends on the task set



- Imagine we try all possible sets



Q: is there a minimum or least upper bound?

A: yes, we can build the task set with least upper bound

Processor Utilization for Rate-Monotonic

- RM priority assignment is optimal
- for a given task set, the U achieved by RM priority assignment is \geq the U for any other priority assignment
- the least upper bound of U = the minimum U_{ub} for RM priority assignment over all possible T 's and all C 's for the tasks

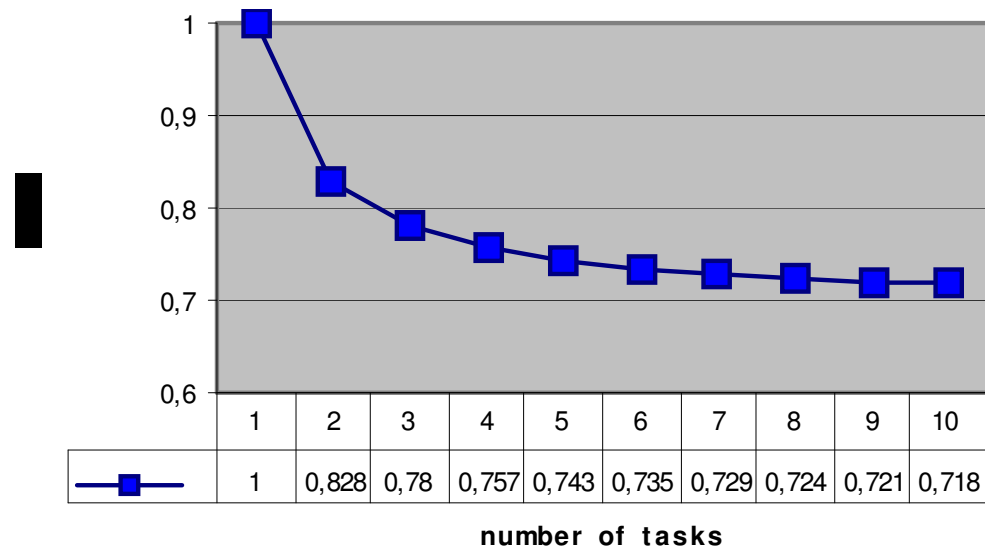
Processor Utilization

- *Theorem: For a set of n tasks with fixed priority assignment, the least upper bound to processor utilization factor is $U=n(2^{1/n}-1)$*
- Or, equivalently, a set of n periodic tasks scheduled by RM algorithm will always meet their deadlines for all task start times if

$$C_1/T_1 + C_2/T_2 + \dots + C_n/T_n \leq n(2^{1/n}-1)$$

Processor Utilization

- If $n \rightarrow \infty$, U converges quickly to $\ln 2 = 0.69$
- sufficient only condition (quite restrictive)
 - what happens to the missing 31%?
- We need a necessary and sufficient condition!



Response time based guarantee

- Response time is the sum of
- Execution time
 - Time spent executing the task
- Non schedulable entities with higher priority
 - Interrupt Handlers
- Scheduling interference
 - Time spent executing higher priority jobs
- Blocking time
 - Time spent executing lower priority tasks
 - Because of access to shared resources
- Applying the critical instant theorem we can compute the worst case completion time (response time) ...

Theorem 1 Recalled

- *Theorem 1: A critical instant for any task occurs whenever the task is requested simultaneously with requests of all higher priority tasks*
- Can use this to determine whether a given priority assignment will yield a feasible scheduling algorithm
 - if requests for all tasks at their critical instants are fulfilled before their respective deadlines, then the scheduling algorithm is feasible
- Applicable to *any* static priority scheme... not just RM

Example #1

- Task τ_1 : $C_1 = 20$; $T_1 = 100$; $D_1 = 100$
Task τ_2 : $C_2 = 30$; $T_2 = 145$; $D_2 = 145$

Is this task set schedulable?

$$U = 20/100 + 30/145 = 0.41 \leq 2(2^{1/2}-1) = 0.828$$

Yes!

Example #2

- Task τ_1 : $C_1 = 20$; $T_1 = 100$; $D_1 = 100$
Task τ_2 : $C_2 = 30$; $T_2 = 145$; $D_2 = 145$
Task τ_3 : $C_3 = 68$; $T_3 = 150$; $D_3 = 150$

Is this task set schedulable?

$$\begin{aligned} U &= 20/100 + 30/145 + 68/150 \\ &= 0.86 > 3(2^{1/3}-1) = 0.779 \end{aligned}$$

Can't say! Need to apply Theorem 1.

Example #2 (contd.)

- Consider the critical instant of τ_3 , the lowest priority task
 - τ_1 and τ_2 must execute at least once before τ_3 can begin executing
 - therefore, completion time of τ_3 is $\geq C_1 + C_2 + C_3 = 20 + 68 + 30 = 118$
 - however, τ_1 is initiated one additional time in $(0, 118)$
 - taking this into consideration, completion time of $\tau_3 = 2 C_1 + C_2 + C_3 = 2 * 20 + 68 + 30 = 138$
- Since $138 < D_3 = 150$, the task set is schedulable

Response Time Analysis for RM

- For the highest priority task, worst case response time R is its own computation time C
 - $R = C$
- Other lower priority tasks suffer interferences from higher priority processes
 - $R_i = C_i + I_i$
 - I_i is the interference in the interval $[t, t+R_i]$

Response Time Analysis (contd.)

- Consider task i , and a higher priority task j
- Interference from task j during R_i :
 - # of releases of task $k = \lceil R_i/T_j \rceil$
 - each will consume C_j units of processor
 - total interference from task $j = \lceil R_i/T_j \rceil * C_j$
- Let $hp(i)$ be the set of tasks with priorities higher than that of task i
- Total interference to task i from all tasks during R_i :

$$I_i = \sum_{j \in hp(i)} \left\lceil \frac{R_i}{T_j} \right\rceil C_j$$

Response Time Analysis (contd.)

- This leads to:

$$R_i = C_i + \sum_{j \in hp(i)} \left[\frac{R_i}{T_j} \right] C_j$$

- Smallest R_i will be the worst case response time
- Fixed point equation: can be solved iteratively

$$w_i^{n+1} = C_i + \sum_{j \in hp(i)} \left[\frac{w_i^n}{T_j} \right] C_j$$

Algorithm

```
for i in 1..N loop -- for each process in turn
  n := 0
   $w_i^n := C_i$ 
  loop
    calculate new  $w_i^{n+1}$  from Equation
    if  $w_i^{n+1} = w_i^n$  then
       $R_i := w_i^n$ 
      exit {value found}
    end if
    if  $w_i^{n+1} > T_i$  then
      exit {value not found}
    end if
    n := n + 1
  end loop
end loop
```

Deadline Monotonic (DM)

- If deadlines are different from the periods, then RM is no more optimal
- If deadlines are lower than periods the Deadline Monotonic policy is optimal among all fixed-priority schemes

Deadline Monotonic (DM)

- Fixed priority of a process is inversely proportional to its deadline (< period)

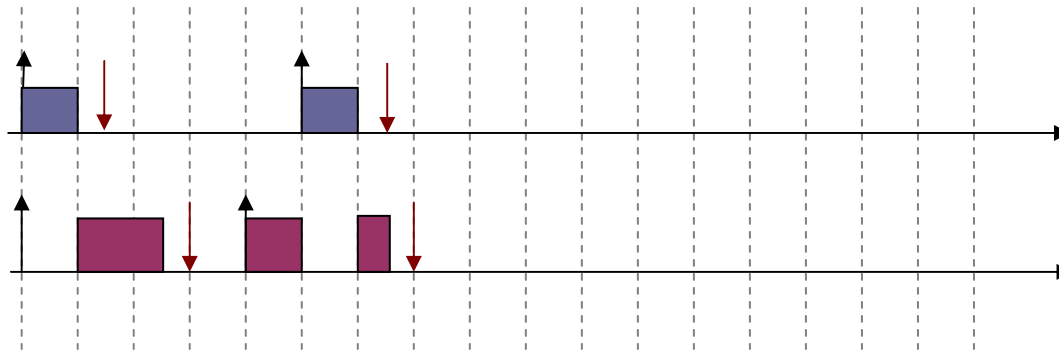
$$D_i < D_j \Rightarrow P_i > P_j$$

- Optimal: can schedule any task set that any other static priority assignment can
- Example: RM fails but DM succeeds for the following

	Period <i>T</i>	Deadline <i>D</i>	Comp Time, <i>C</i>	Priority <i>P</i>	Response Time, <i>R</i>
Task_1	20	5	3	4	3
Task_2	15	7	3	3	6
Task_3	10	10	4	2	10
Task_4	20	20	3	1	20

Deadline Monotonic (DM)

- The sufficient-only utilization bound is very pessimistic ...



- The set (C, D, T) $\tau_1=(1, 1.5, 5)$ and $\tau_2=(1.5, 3, 4)$ is schedulable even if ...

$$\sum_i C_i/D_i = 1/1.5 + 1.5/3 = 0.66 + 0.5 = 1.16 > 1$$

Can one do better?

- Yes... by using dynamic priority assignment
- In fact, there is a scheme for dynamic priority assignment for which the least upper bound on the processor utilization is 1
- More later...


Arbitrary Deadlines

- Case when deadline $D_i < T_i$ is easy...
- Case when deadline $D_i > T_i$ is much harder
 - multiple iterations of the same task may be alive simultaneously
 - may have to check multiple task initiations to obtain the worst case response time
- Example: consider two tasks
 - Task 1: $C1 = 28, T1 = 80$
 - Task 2: $C2 = 71, T2 = 110$
 - Assume all deadlines to be infinity

Arbitrary Deadlines (contd.)

- Response time for task 2:

activation	completion time	response time
0	127	127
110	226	116
220	353	133
330	452	122
440	551	111
550	678	128
660	777	117
770	876	106



- Response time is worst for the third instance (not the first one at the critical instant !)
 - Not sufficient to consider just the first iteration

Arbitrary Deadlines (contd.)

- Furthermore, deadline monotonic priority assignment is not optimal anymore ...
- Let $n = 2$ with
- $C_1 = 52, T_1 = 100, D_1 = 110$
- $C_2 = 52, T_2 = 140, D_2 = 154.$
- if τ_1 has highest priority, the set is not schedulable (first instance of τ_2 misses its deadline)
- if τ_2 has highest priority ...

t1 response times

104

208

260

t2 response times

52

192

332

Arbitrary Deadlines (contd.)

- Can we find a schedulability test ?
 - Yes
- Can we find an optimal priority assignment ?
 - Yes

Schedulability Condition for Arbitrary Deadlines

- Analysis when D_i (and hence potentially R_i) can be greater than T_i

$$w_i^{n+1}(q) = (q+1)C_i + \sum_{j \in hp(i)} \left\lceil \frac{w_i^n(q)}{T_j} \right\rceil C_j$$

$$R_i(q) = w_i(q) - qT_i$$

- The number of releases that need to be considered is bounded by the lowest value q^* of $q = 0, 1, 2, \dots$ for which the following relation is true:

$$q^* = \min_q R_i(q) \leq T_i$$

- Note: for $D \leq T$, the condition is true for $q=0$ if the task can be scheduled, in which case the analysis simplifies to original
 - if any $R > D$, the task is not schedulable

Arbitrary Deadlines (contd.)

- The worst-case response time is then the maximum value found for each q :

$$R_i = \max_{q=0, \dots, q^*} R_i(q)$$

Optimal priority assignment for Arbitrary Deadlines

- Audsley's algorithm

```
PriorityAssignment( $\Delta$ )
{
  for j in (n..1) {
    unassigned = TRUE
    for  $\tau_A$  in  $\Delta$  {
      if ((feasible( $\tau_A$ , j)) {
         $\Psi(j) = \tau_A$ 
         $\Delta = \Delta - \tau_A$ 
        unassigned = FALSE
      }
    }
    if (unassigned)
      exit // NOT SCHEDULABLE
  }
}
```

Glossary

Δ	set of all tasks
j	priority level
feasible()	feasibility test
$\Psi(j)$	inverse of priority level assignment function

- Processor demand criterion

Response Time Analysis

- Response time Analysis runs in pseudopolynomial time
- Is it possible to know a-priori the time intervals over which the test should be performed?
 - The iterative procedure tests against increasing intervals corresponding to the w_i^k
- The alternative method is called **processor demand criterion**
- It applies to the case of static and dynamic priority

Fixed Priority Scheduling

- Utilization-based Analysis
- Response time Analysis
- Processor Demand Analysis
 - Important: allows for sensitivity analysis

Processor Demand Analysis

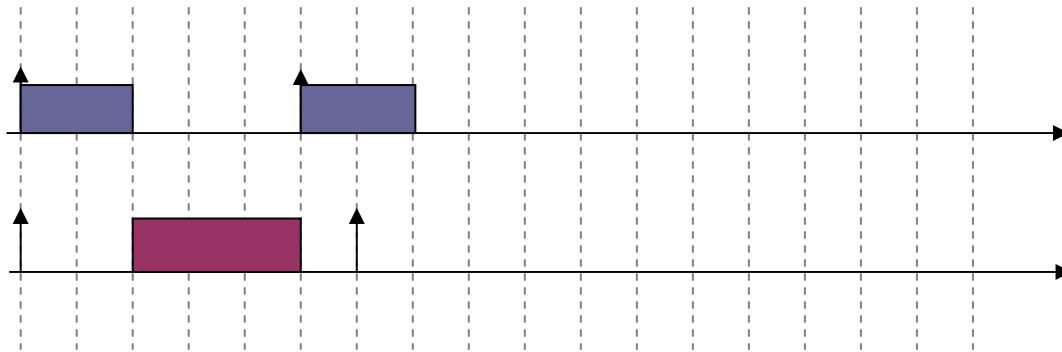
- Consider tasks $\tau_1, \tau_2, \dots, \tau_n$ in decreasing order of priority
- For task τ_i to be schedulable, a necessary and sufficient condition is that we can find some $t \in [0, T_i]$ satisfying the condition

$$t = \lceil t/T_1 \rceil C_1 + \lceil t/T_2 \rceil C_2 + \dots + \lceil t/T_{i-1} \rceil C_{i-1} + C_i$$

- But do we need to check at exhaustively for all values of t in $[0, T_i]$?

Processor Demand Analysis

- Clearly only T_i is not enough ...
- Example: consider the set $\tau_1=(2, 5)$ and $\tau_2=(3,6)$
- The processor demand for τ_2 in $[0,6]$ is 7 units
- ... but the system is clearly schedulable since the processor demand in $[0,5]$ is 5 units



Processor Demand Analysis

- Observation: right hand side of the equation changes only at multiples of T_1, T_2, \dots, T_{i-1}
- It is therefore sufficient to check if the inequality is satisfied for some $t \in [0, T_i]$ that is a multiple of one or more of T_1, T_2, \dots, T_{i-1}

$$t \geq \lceil t/T_1 \rceil C_1 + \lceil t/T_2 \rceil C_2 + \dots + \lceil t/T_{i-1} \rceil C_{i-1} + C_i$$

Processor Demand Analysis

- Notation

$$W_i(t) = \sum_{j=1..i} C_j \lceil t/T_j \rceil$$

$$L_i(t) = W_i(t)/t$$

$$L_i = \min_{0 \leq t \leq T_i} L_i(t)$$

$$L = \max\{L_i\}$$

- General sufficient & necessary condition:

- Task τ_i can be scheduled iff $L_i \leq 1$

- Practically, we only need to compute $W_i(t)$ at all times

$$\alpha_i = \{kT_j \mid j=1, \dots, I; k=1, \dots, \lfloor T_i/T_j \rfloor\}$$

- these are the times at which tasks are released

- $W_i(t)$ is constant at other times

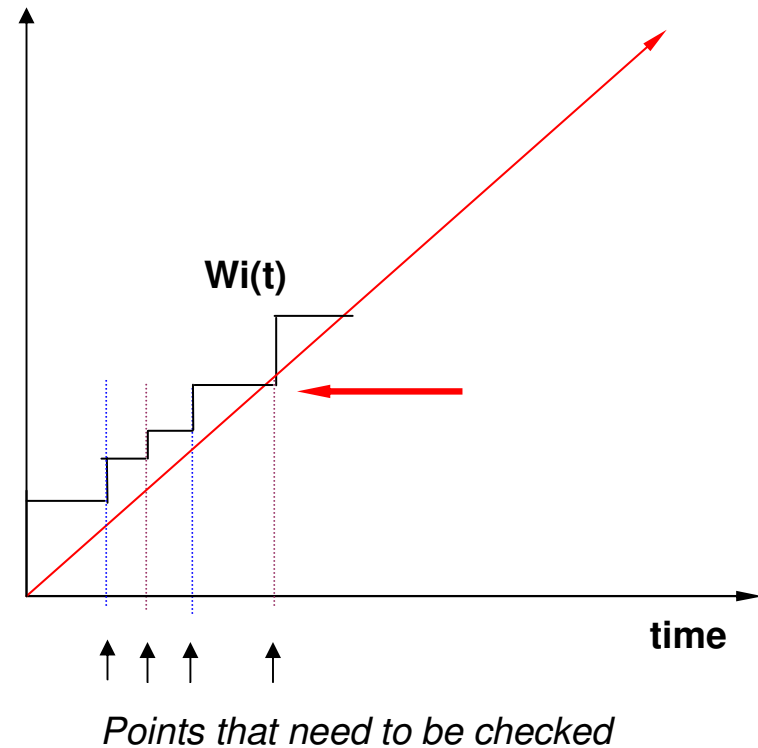
- Practical RM schedulability conditions:

- if $\min_{t \in \alpha_i} W_i(t)/t \leq 1$, task τ_i is schedulable

- if $\max_{i \in \{1, \dots, n\}} \{\min_{t \in \alpha_i} W_i(t)/t\} \leq 1$, then the entire set is schedulable

Example

- Task set:
 - $\tau_1: T_1=100, C_1=20$
 - $\tau_2: T_2=150, C_2=30$
 - $\tau_3: T_3=210, C_3=80$
 - $\tau_4: T_4=400, C_4=100$
- Then:
 - $\alpha_1 = \{100\}$
 - $\alpha_2 = \{100, 150\}$
 - $\alpha_3 = \{100, 150, 200, 210\}$
 - $\alpha_4 = \{100, 150, 200, 210, 300, 400\}$



- Plots of $W_i(t)$: task τ_i is RM-schedulable iff any part of the plot of $W_i(t)$ falls on or below the $W_i(t)=t$ line.
- We will improve this formulation (see next slide(s) ...)

Processor Demand Analysis

- Improvement [Bini]

Theorem 1 (Theorem 3 in [2]) A task set $\mathcal{T} = \{\tau_1, \tau_2, \dots, \tau_n\}$ is schedulable *if and only if*:

$$\forall i = 1 \dots n \quad \bigvee_{t \in \mathcal{P}_{i-1}(T_i)} \sum_{j=1}^i \left\lceil \frac{t}{T_j} \right\rceil C_j \leq t \quad (2)$$

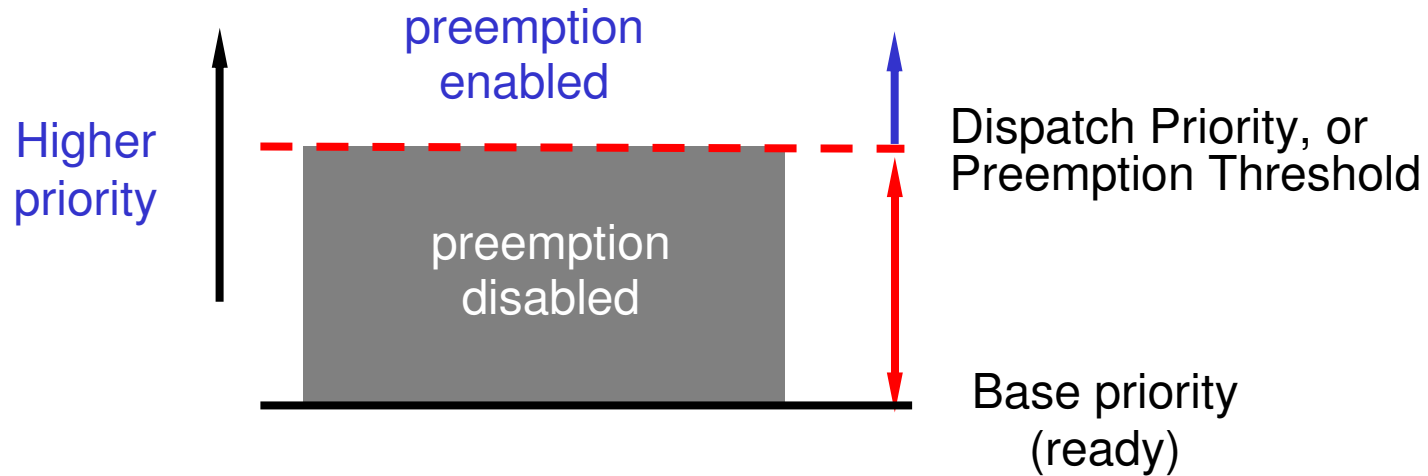
where $\mathcal{P}_i(t)$ is defined by the following recurrent expression:

$$\begin{cases} \mathcal{P}_0(t) = \{t\} \\ \mathcal{P}_i(t) = \mathcal{P}_{i-1} \left(\left\lfloor \frac{t}{T_i} \right\rfloor T_i \right) \cup \mathcal{P}_{i-1}(t). \end{cases} \quad (3)$$

Preemption Threshold (dual priority)

- Derived from Fixed priority scheduling theory
 - Also available for dynamic priority policies
- Uses two priority levels for each task
 - **Ready** priority for enqueueing tasks in the ready queue
 - **Dispatch** priority for preempting the currently executing task
 - Ready priority \leq Dispatch priority
- Advantages
 - May perform better than purely preemptive or non-preemptive schemes
 - Allows selectively disabling preemption

Preemption Threshold (dual priority)



- Preemptive Scheduling
 - Dispatch Priority = Base Priority
- Non-Preemptive Scheduling
 - Dispatch Priority = Maximum Priority

Preemption Threshold (dual priority) : esempio

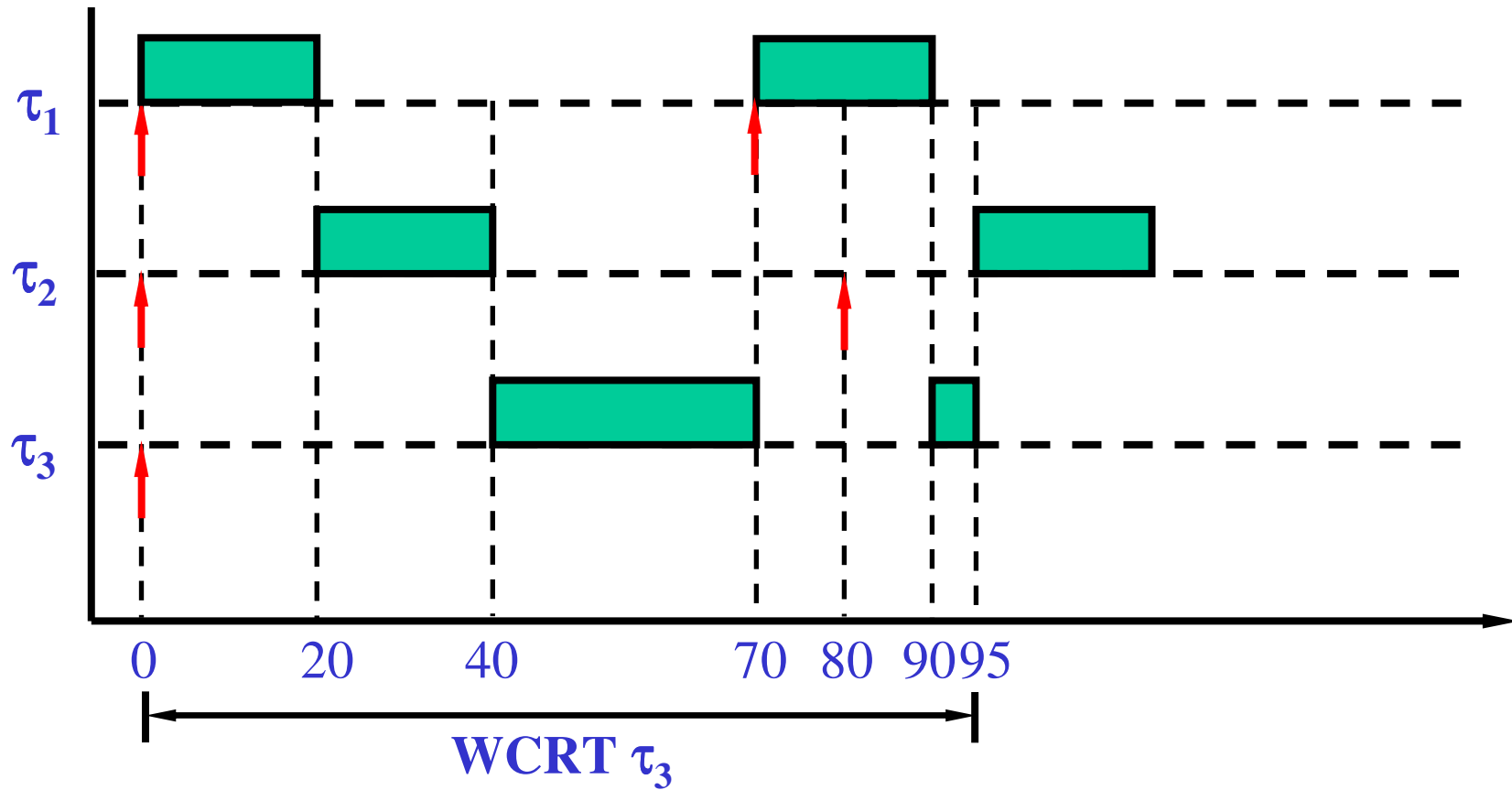
- Task

Tasks	C_i	T_i	D_i	π_i
τ_1	20	70	50	1
τ_2	20	80	80	2
τ_3	35	200	100	3

- Worst-case Response Time

Tasks	π_i	γ_i	WCRT <i>Preemptive</i>	WCRT <i>Non-Preemptive</i>	WCRT <i>With Threshold</i>
τ_1	1	1	20	55	40
τ_2	2	1	40	75	75
τ_3	3	2	115	75	95

Preemption Threshold (dual priority): an example



Preemption Threshold: analysis

- Before a task τ_i starts execution, there is blocking from lower priority tasks and interference from higher priority tasks. Among all lower priority tasks, only one lower priority task can cause blocking. The maximum blocking time of a task τ_i , denoted by $B(\tau_i)$, is given by:

$$B(\tau_i) = \max_{\forall j, \gamma_j \geq \pi_i > \pi_j} C_j$$

- All higher priority tasks that come before the start time $S_i(q)$ and any earlier instances of task τ_i before instance q should be finished before the q -th start time.

$$S_i(q) = B(\tau_i) + (q - 1) \cdot C_i + \sum_{\forall j, \pi_j > \pi_i} \left(1 + \left\lfloor \frac{S_i(q)}{T_j} \right\rfloor \right) \cdot C_j$$

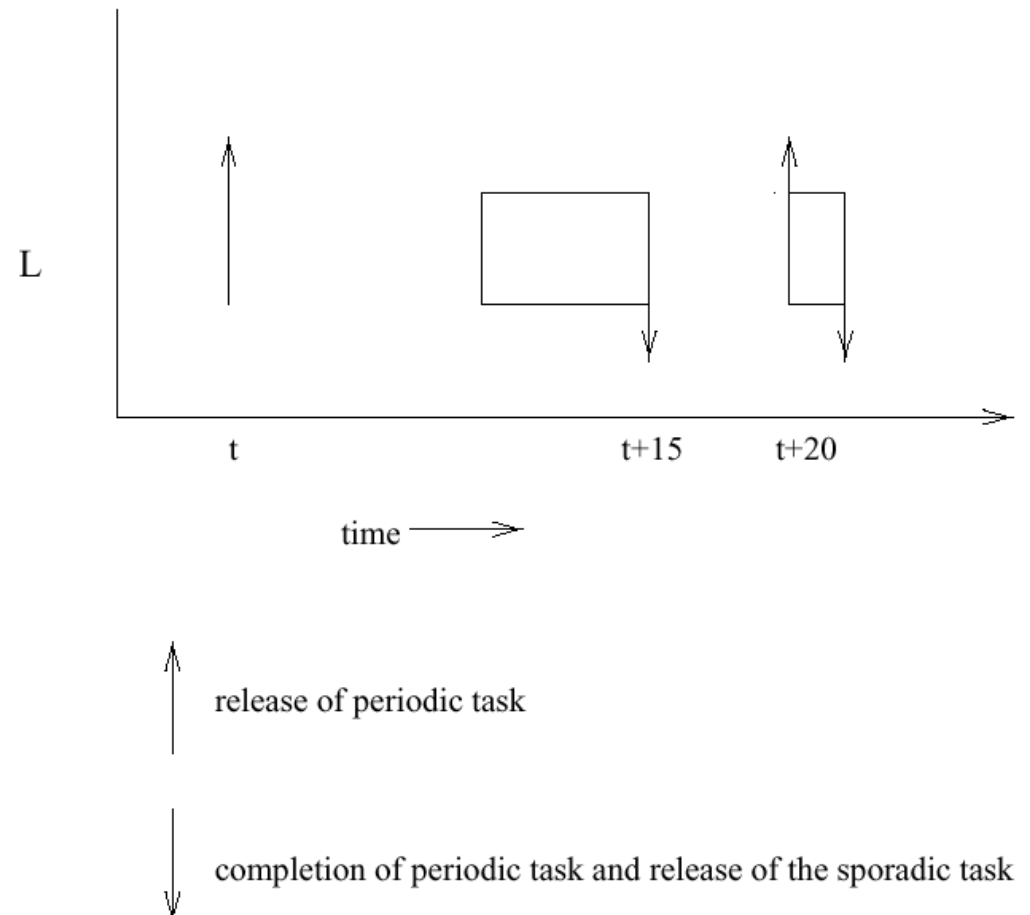
Preemption Threshold: analysis

- Once the q -th instance of task τ_i starts execution, we have to consider the interference to compute its finish time. From the definition of preemption threshold, we know that only tasks with higher priority than the preemption threshold of τ_i can preempt τ_i and get the CPU before it finishes. Furthermore, we only need to consider new arrivals of these tasks, i.e., arrivals after $S_i(q)$.

$$\mathcal{F}_i(q) = S_i(q) + C_i + \sum_{\forall j, \pi_j > \gamma_i} \left(\left\lceil \frac{\mathcal{F}_i(q)}{T_j} \right\rceil - \left(1 + \left\lfloor \frac{S_i(q)}{T_j} \right\rfloor \right) \right) \cdot C_j$$

Release Jitter

- A key issue in distributed systems
- Sporadic task will be released at time 0, 5, 25, 45, and so on ...
- i.e. at times 0, $T-J$, $2T-J$, $3T-J$, and so on...



Release Jitter (contd.)

- Examination of the derivation of the schedulability equation implies that process i will suffer one interference from S if R_i is between 0 and $T-J$, that is $R_i \in [0, T-J)$, two if $R_i \in [T-J, 2T-J)$, three if $R_i \in [2T-J, 3T-J)$, and so on...

$$R_i = B_i + C_i + \sum_{j \in hp(i)} \left\lceil \frac{R_i + J_j}{T_j} \right\rceil C_j$$

Release Jitter (contd.)

- In general, periodic tasks do not suffer jitter
- But, an implementation may restrict granularity of system timer which releases periodic tasks
 - a periodic task may therefore suffer from jitter
- If response time is to be measured relative to the real release time then the jitter value must be added to that previously calculated:

$$R_i^{\text{periodic}} = R_i + J_i$$

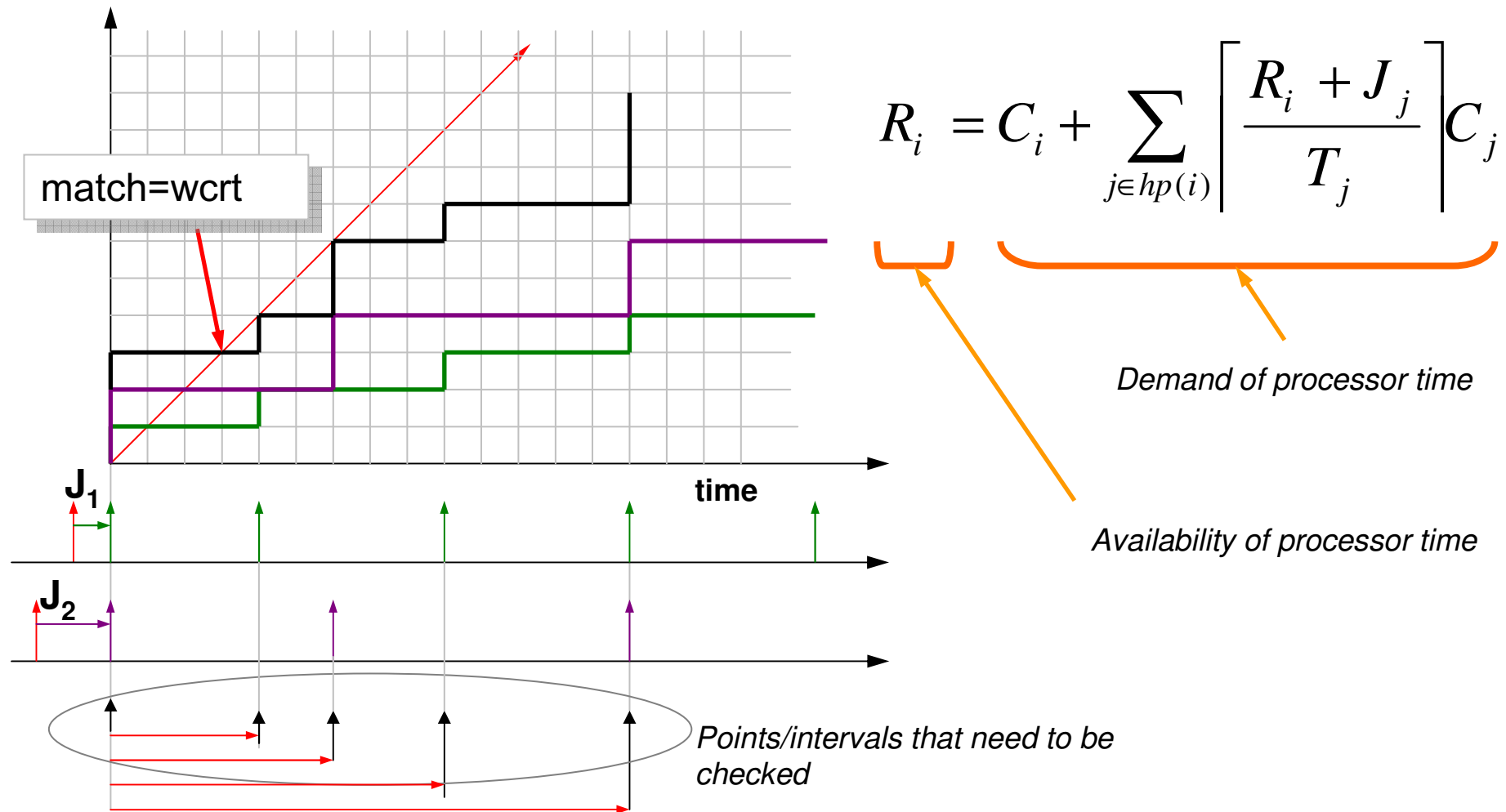
Arbitrary Deadlines with Release Jitter

$$w_i^{n+1}(q) = B_i + (q+1)C_i + \sum_{j \in hp(i)} \left\lceil \frac{w_i^n(q) + J_j}{T_j} \right\rceil C_j$$

$$R_i(q) = w_i^n(q) - qT_i + J_i$$

Tasks with Jitter/Processor demand (dbf)

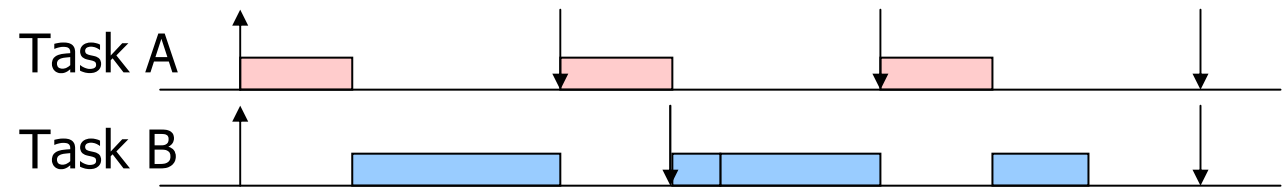
- Task τ_1 : $T_1=50$, $C_1=10$ $J_1=10$, τ_2 : $T_2=80$, $C_2=20$ $J_2=20$



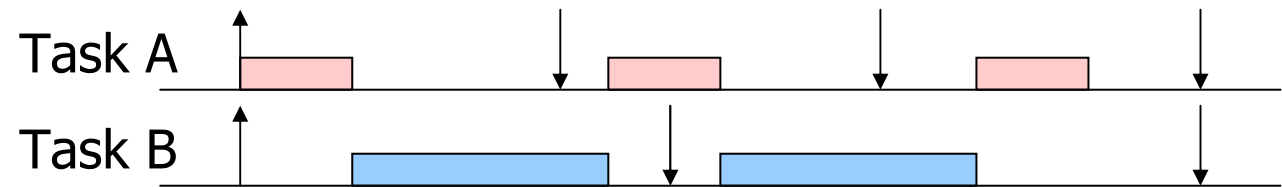
Earliest Deadline First

- With EDF (dynamic priorities) the utilization bound is 100%

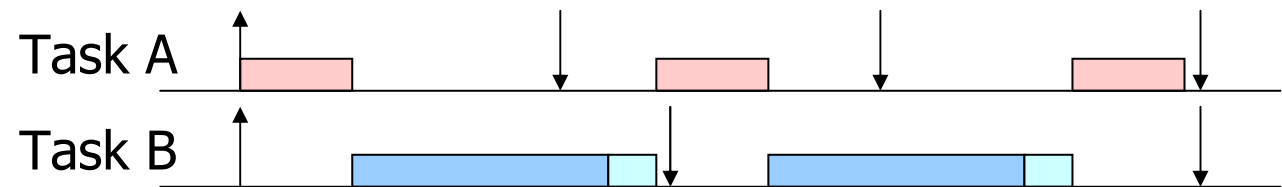
- RM



- EDF...

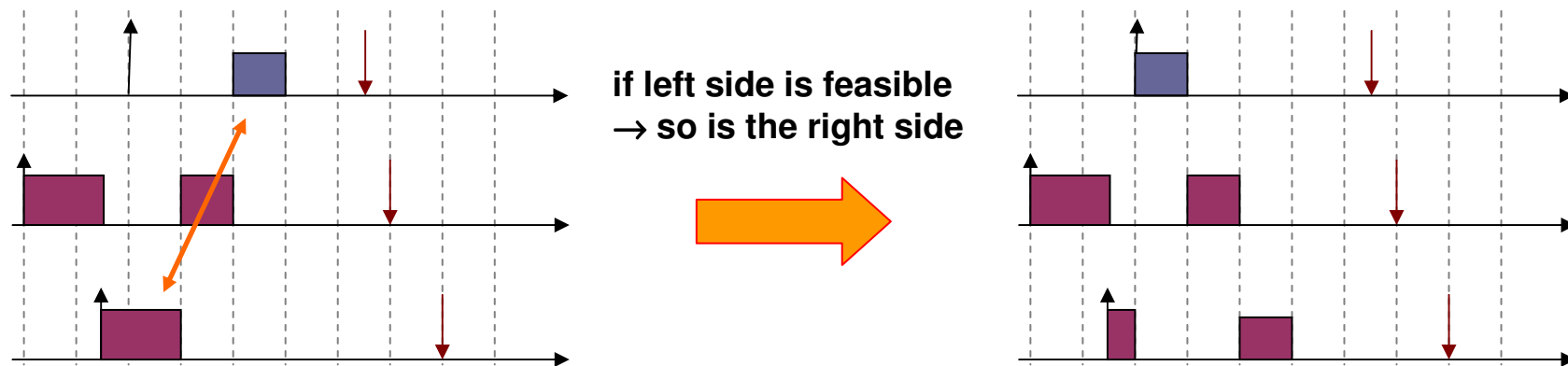


...utilization
can be
further
increased !



Earliest Deadline First

- EDF is clearly optimal among all scheduling schemes
- Proof for any D: interchange argument [Dertouzos '74]



- (Proof $D=T$: [LiuLayland73] follows from utilization bound=100%)

Earliest Deadline First



- There are few (if any) commercial implementations of EDF

“EDF implementations are inefficient and should be avoided because a RT system should be as fast as possible”

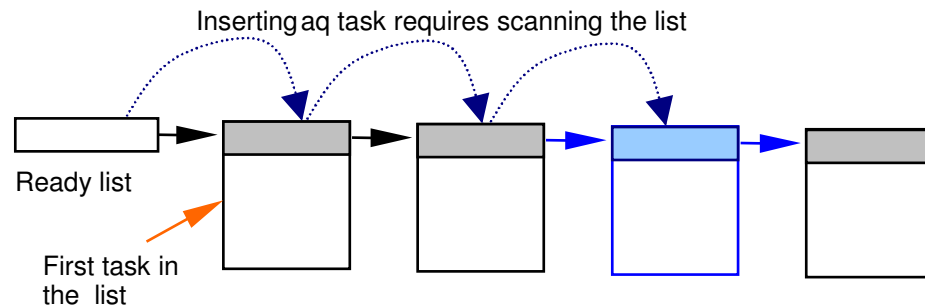
“EDF cannot be controlled in overload conditions”

Implementation of Earliest Deadline First

- Is it really not feasible to implement EDF scheduling ?
- Problems
 - absolute deadlines change for each new task instance, therefore the priority needs to be updated every time the task moves back to the ready queue
 - more important, absolute deadlines are always increasing, how can we associate a (finite) priority value to an ever-increasing deadline value
 - most important, absolute deadlines are impossible to compute a-priori (there are infinitely many). Do we need infinitely many priority levels?
 - What happens in overload conditions?

Implementation of fixed priority

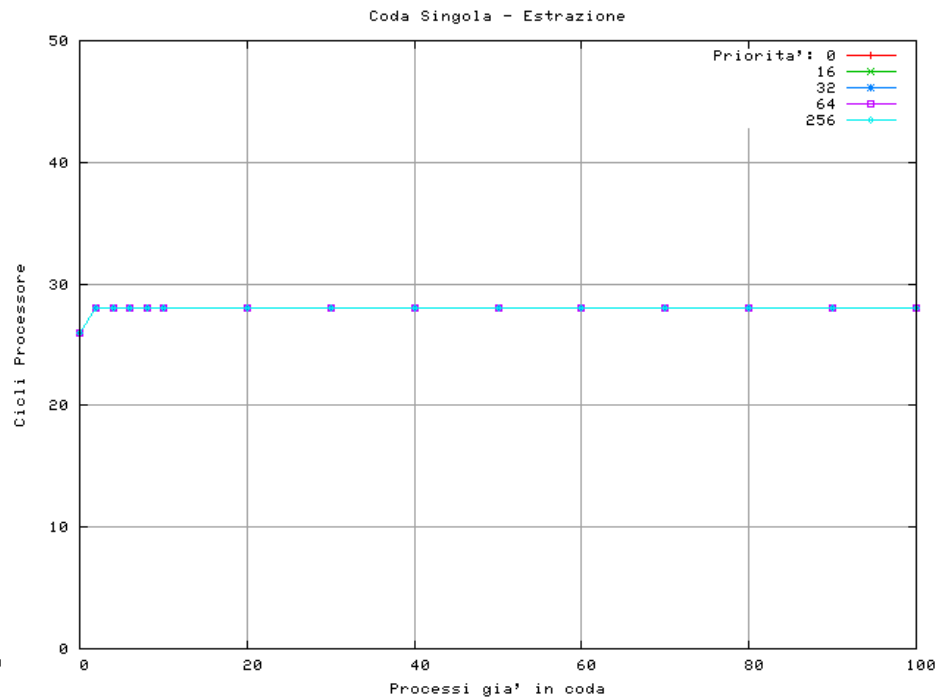
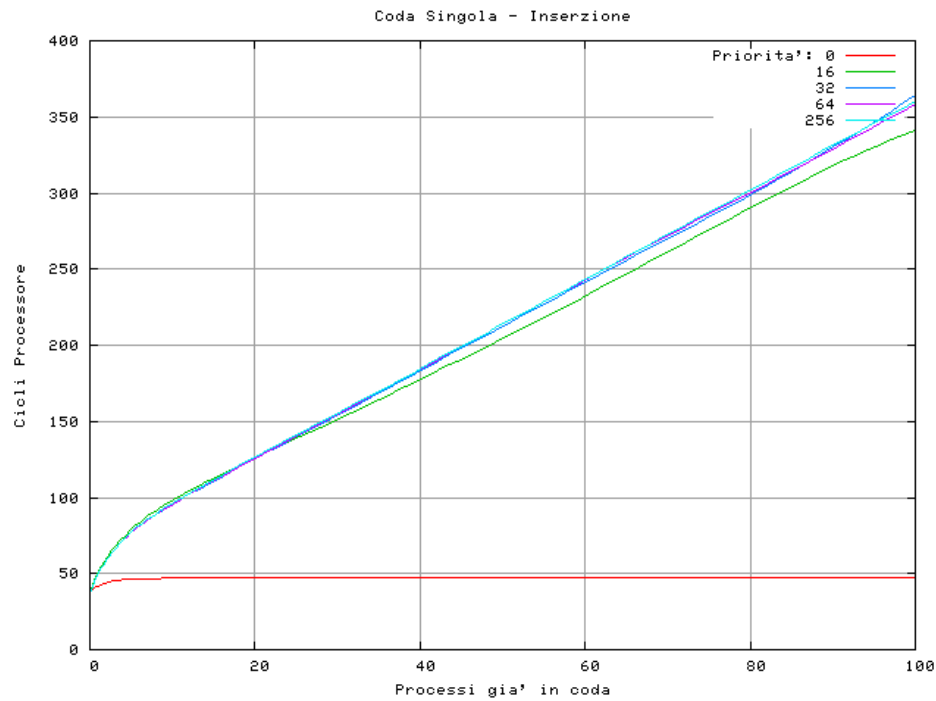
- When implementing fixed priority scheduling, it is possible to build ready queues and semaphore queues with constant-time insertion and extraction times (at the price of some memory)



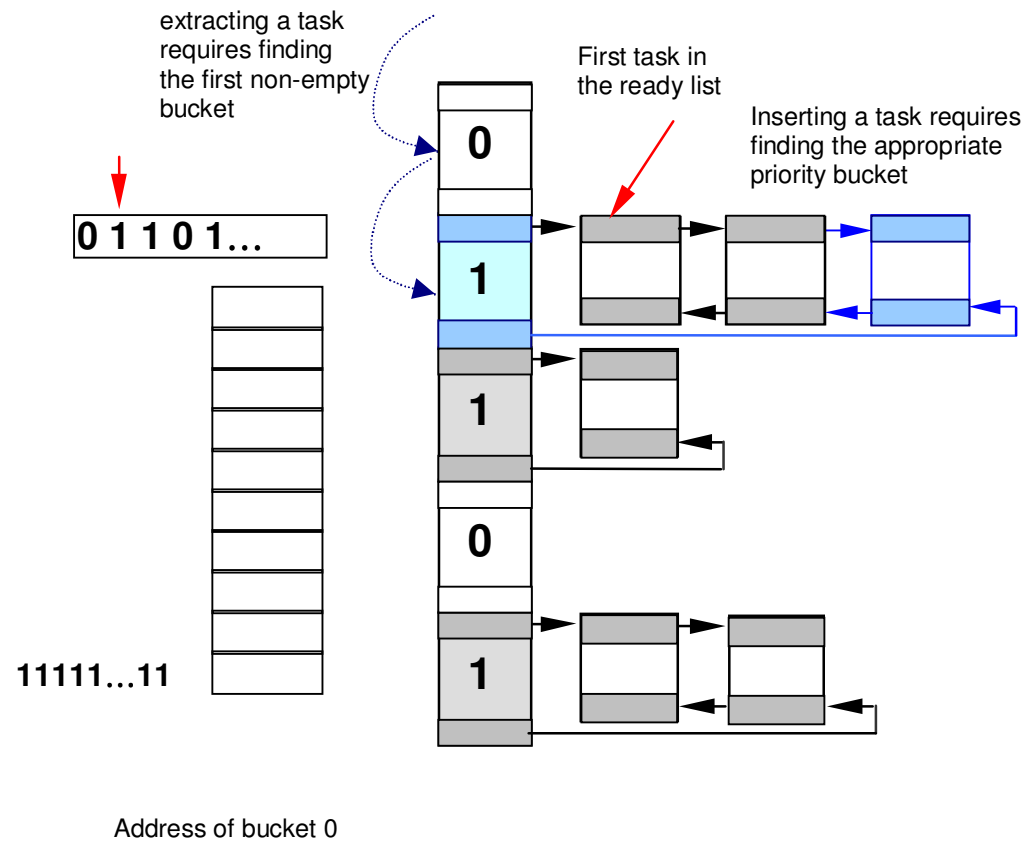
Step 1: simple ready list
(extraction $O(1)$ insertion $O(n)$) where n is the number of task descriptors

Implementation of fixed priority

- Simple queue experimental measures



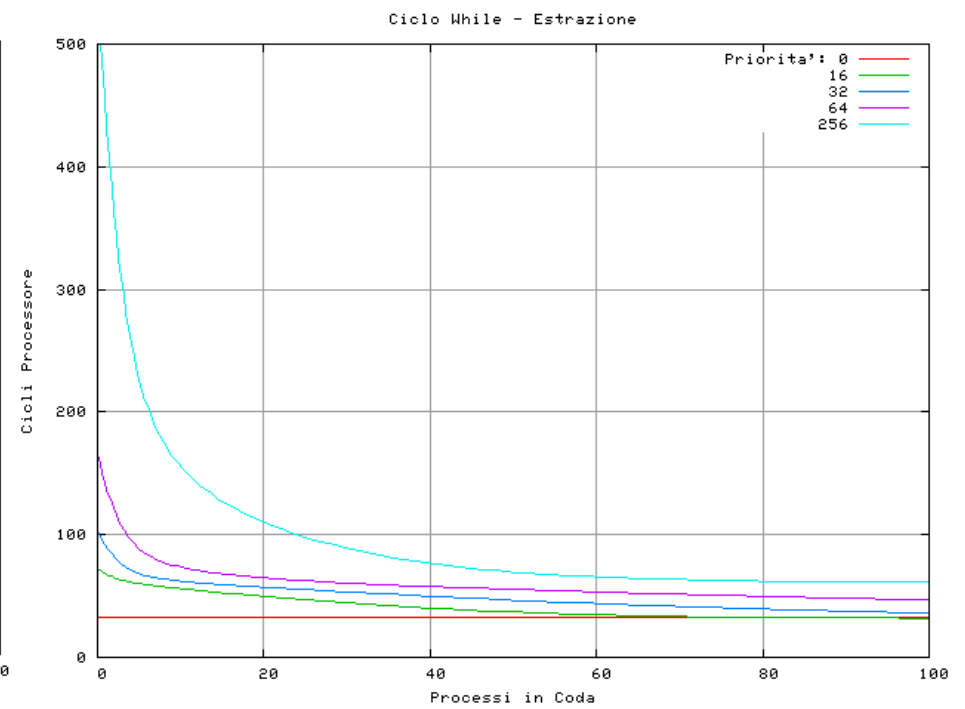
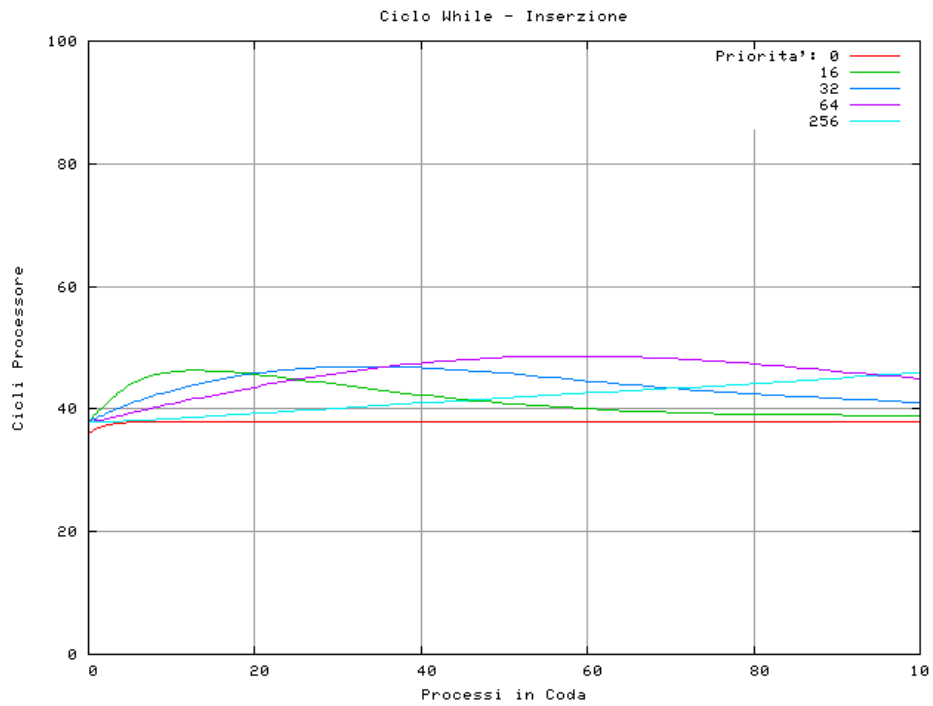
Implementation of fixed priority



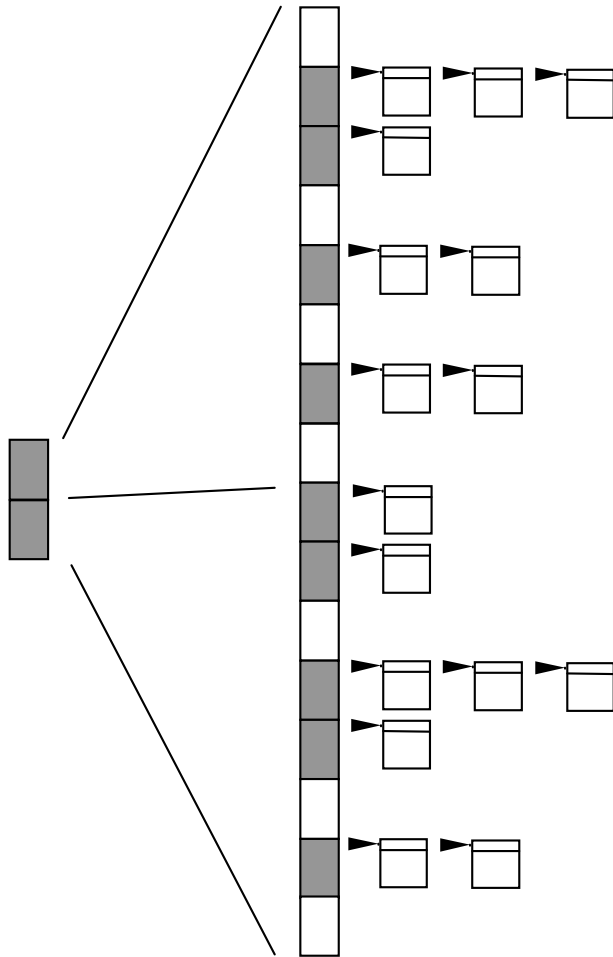
Step 2: multiple priority queue ready list (extraction $O(m)$ insertion $O(1)$) where m is the number of priorities

Implementation of fixed priority

- Multiple queue experimental measures



Implementation of fixed priority

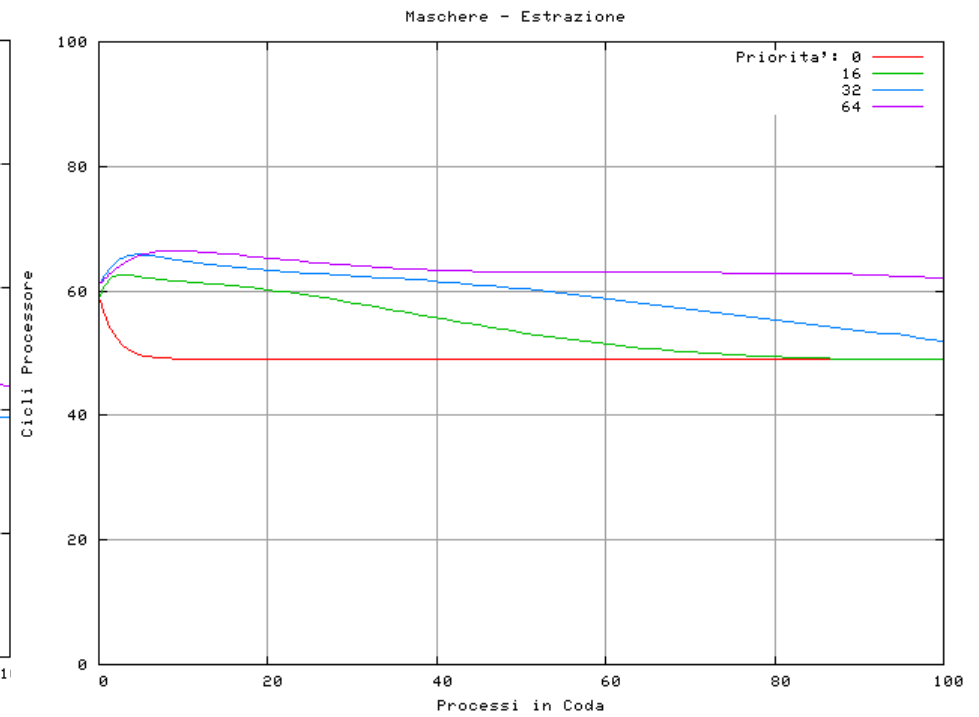
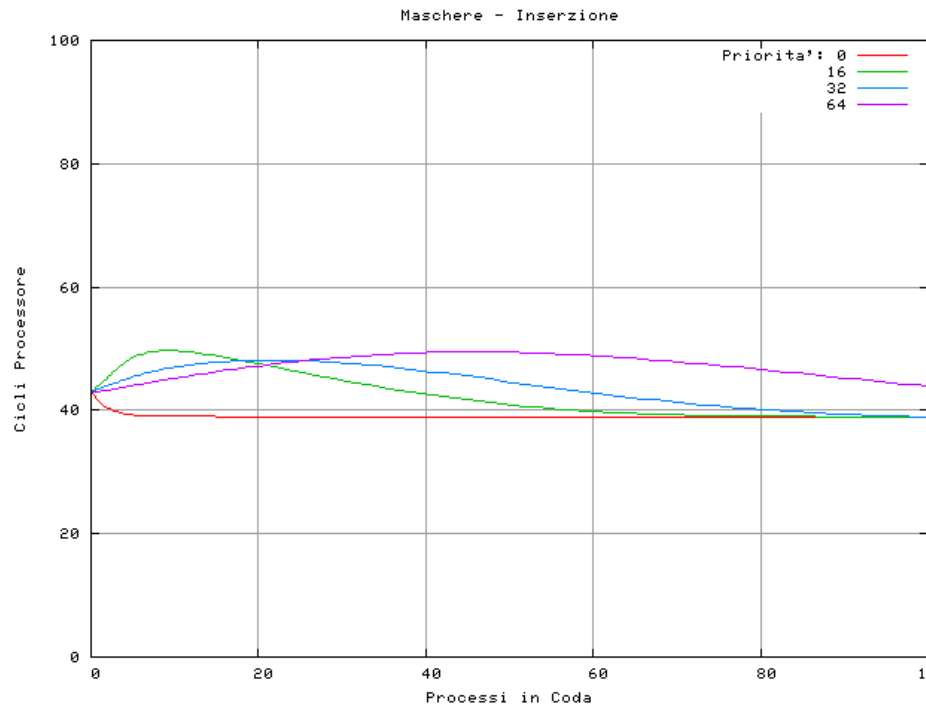


Step 3: hierarchical priority queues (extraction $O(\log_b m)$ insertion $O(1)$) + lookup tables in order to avoid bit shifting. B is the size (in bits) of the bitmask containing the status of the priority queues

If $m=256$ and $b=8$ than extraction is in constant time (2 steps)

Implementation of fixed priority

- Bitmapped queue experimental measures



Count Leading Zeros (where available)

ARM Architecture Reference Manual

The ARM Instruction Set

3.6 Miscellaneous arithmetic instructions

In addition to the normal data-processing and multiply instructions, versions 5 and above of the ARM architecture include a Count Leading Zeros (CLZ) instruction. This instruction returns the number of 0 bits at the most significant end of its operand before the first 1 bit is encountered (or 32 if its operand is zero). Two typical applications for this are:

- To determine how many bits the operand should be shifted left in order to *normalize* it, so that its most significant bit is 1. (This can be used in integer division routines.)
- To locate the highest priority bit in a bit mask.

3.6.1 Instruction encoding

CLZ{<cond>} <Rd>, <Rm>

31	28	27	26	25	24	23	22	21	20	19	16	15	12	11	8	7	6	5	4	3	0
cond		0 0 0 1 0 1 1 0						SBO		Rd		SBO		0 0 0 1			Rm				

Rd Specifies the destination register.

Rm Specifies the operand register.

3.6.2 List of miscellaneous arithmetic instructions

CLZ Count Leading Zeros. See *CLZ* on page A4-22.

Implementation of fixed priority

From an evaluation of VxWorks 5.3 (www.embedded-systems.com)

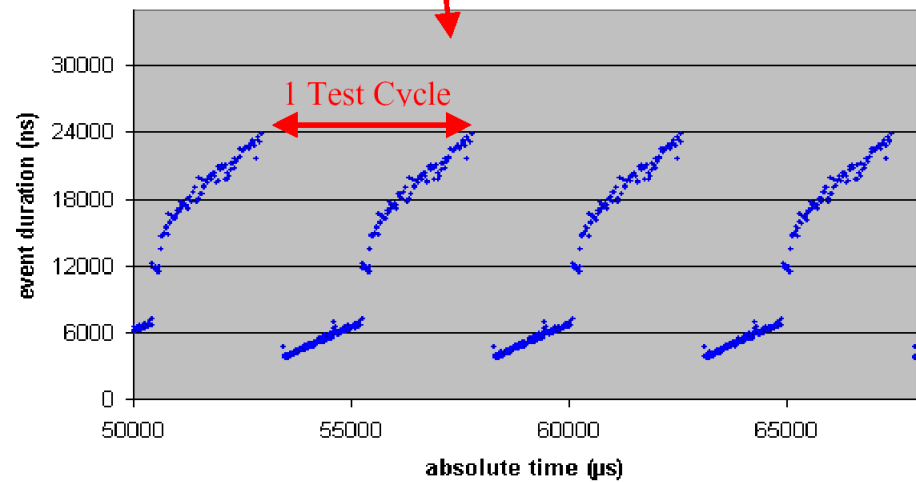
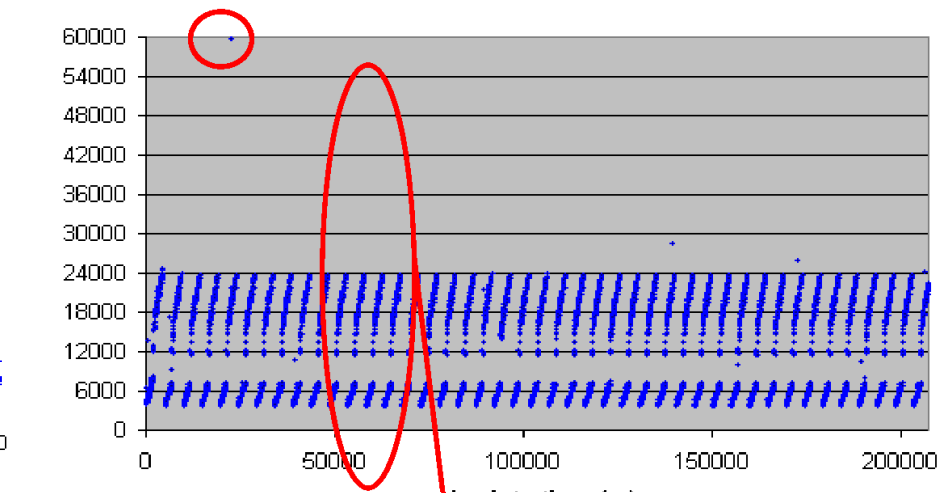
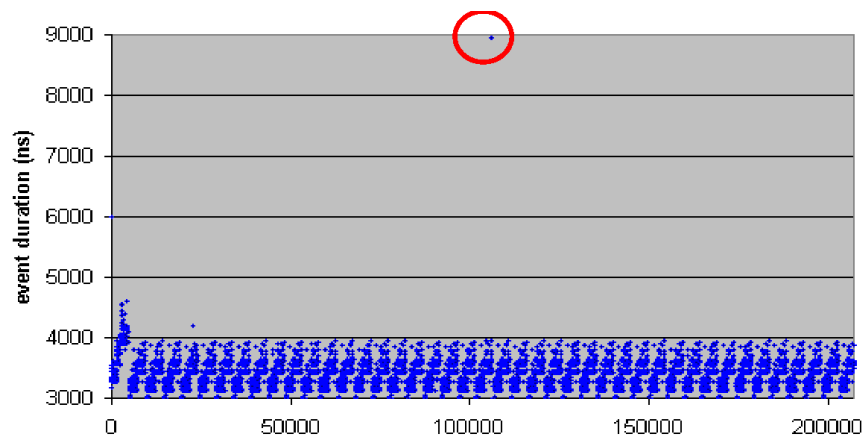
In this test, we measure the time it takes the system to release a binary semaphore and schedule a higher priority thread that was thereby released. A number of threads is increased one by one until there are 255 threads of different priority pending on the

The idea is to investigate whether or not the time to release the semaphore (and schedule the released thread) is proportional with the number of threads waiting for the semaphore.

However, we reprocessed the test results to find out how long it takes for a thread to acquire a semaphore that is not available. When a thread acquires a semaphore that is not available, the thread needs to be added to the semaphore's queue of waiting threads. Good RTOS design requires this queue to be sorted at all times in order to keep the release time of a semaphore constant (as described in the previous paragraph). The structure of this queue is therefore very important, in order to keep the sorting time as constant and as short as possible. This did not happen in VxWorks 5.3.1 as can be seen from Figure 4.7-13 and Figure 4.7-14 (p.57). It is clear that the time it takes to add a thread to the semaphore's queue (and sort it) is proportional to the number of threads already in the queue. Queue structures that lead to better (and more constant) sorting latencies are available but require more memory, which may be unacceptable for systems with tight memory constraints.

Implementation of fixed priority

From an evaluation of VxWorks 5.3



Implementation of Earliest Deadline First

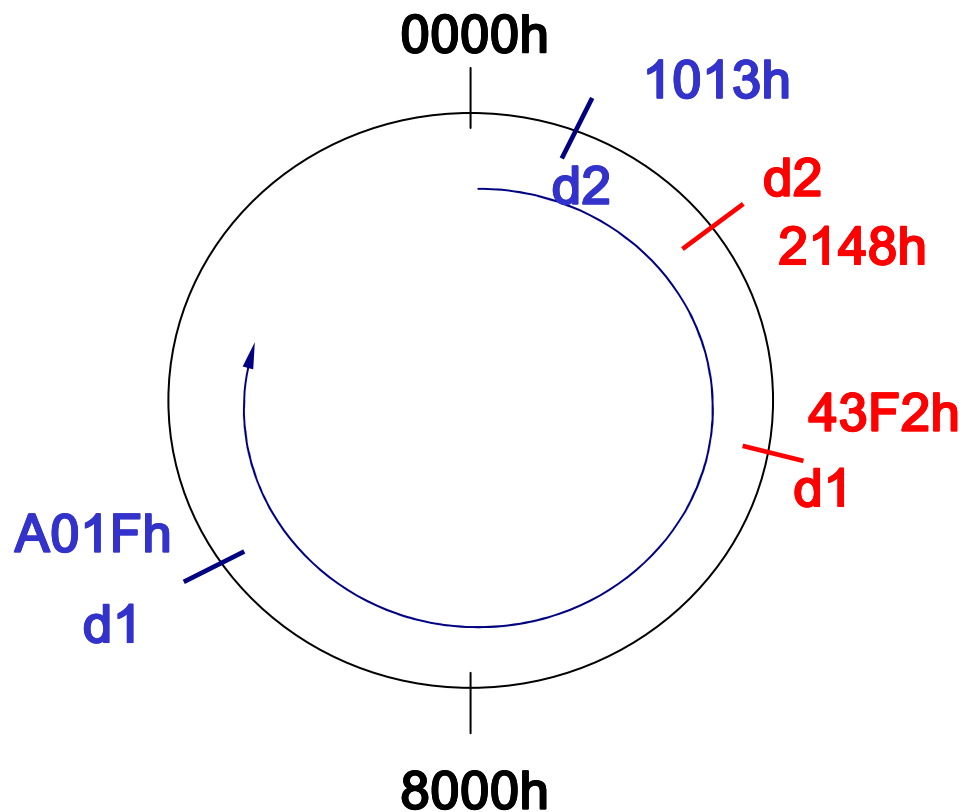
- Problem 2: deadline encoding ?
 - The EDF scheduler, requires a time reference to compute the absolute deadline (the priority) of a newly activated task. Such a timer must necessarily feature a long lifetime and a short granularity. For example, in POSIX systems, a 64 bit structure allows for a granularity of nanoseconds. In an embedded system, such a high precision might actually become undesirable since it leads to an unacceptable overhead.

Implementation of Earliest Deadline First

- Problem 2: deadline encoding ?
 - The problem can be efficiently solved using a limited resolution (i.e. 16 bit) timer and an algorithm first described in [Fonseca01]. Suppose the current timer value and the absolute deadlines are represented as 16 bit words. Each time a task is activated, the system computes an absolute deadline for it as the current timer value plus the task's relative deadline: this operation could result in an overflow. However, ignoring overflows, it is still possible to compare two absolute deadlines in a consistent way. Suppose that the maximum relative deadline is less than $7FFFh$ timer ticks, and let δ be the difference between two absolute deadlines d_1 and d_2 : δ is always in the interval $[-8000h; +7FFFh]$ and can be expressed as a signed 16 bit integer. The sign of δ can be used as a way to compare d_1 and d_2 : if $\delta > 0$ then $d_1 > d_2$.

Implementation of Earliest Deadline First

- This compare algorithm is very simple and can efficiently be implemented with two simple operations: a difference between integers and a sign check.



$$d1-d2 = 43F2h-2148h = 22AA > 0$$

↓
d1>d2

$$d1-d2 = A01Fh-1013h = 900C < 0$$

↓
d1<d2

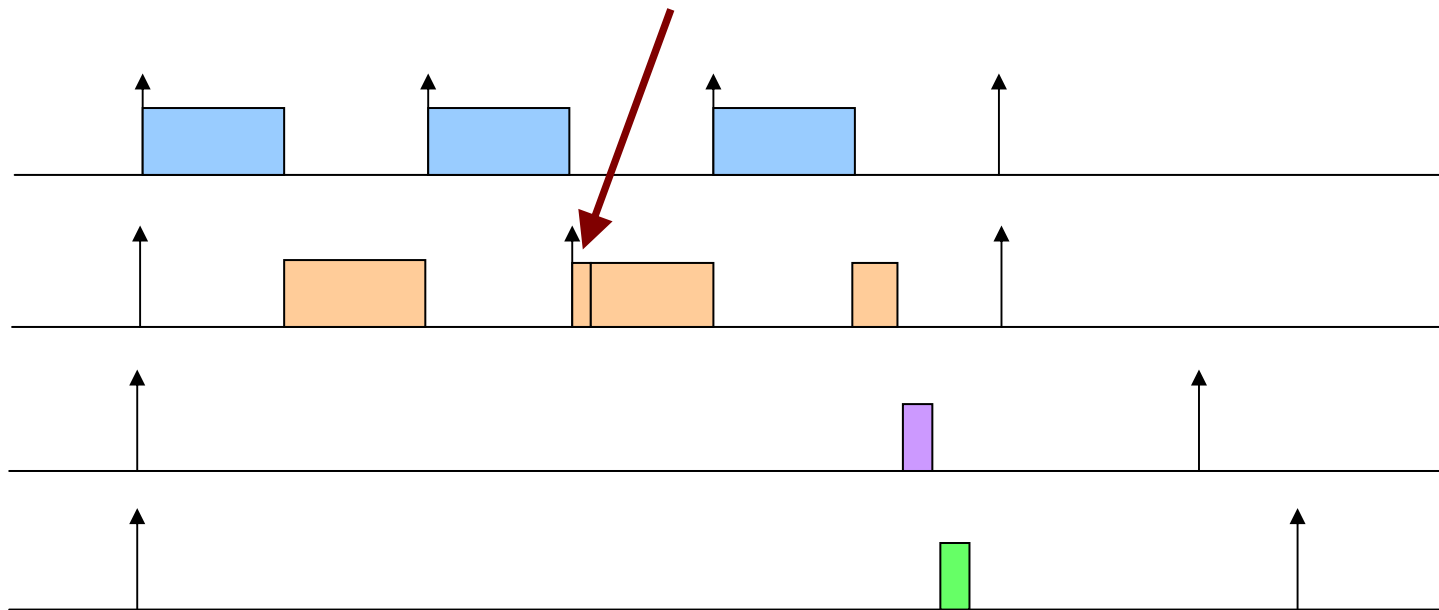
Implementation of Earliest Deadline First

- Overload conditions
- EDF can give rise to a cascade of deadline miss
 - There is no guarantee on which is the task that will miss its deadline
 - (see also problems with determination of worst case completion time)
- Try the case
 - $C_1=1$ $T_1=4$
 - $C_2=2$ $T_2=6$
 - $C_3=2$ $T_3=8$
 - $C_4=3$ $T_4=10$

(utilization = 106%)

Overload in FP scheduling

- Overload conditions
- Misconception: In FP the lowest priority tasks are the first to miss the deadline
- Counterexample: start from the set (2,4) (2,6) fully utilizing the processor



Task Synchronization

- So far, we considered independent tasks
- However, tasks do interact: semaphores, locks, monitors, rendezvous etc.
 - shared data, use of non-preemptable resources
- This jeopardizes systems ability to meet timing constraints
 - e.g. may lead to an indefinite period of *priority inversion* where a high priority task is prevented from executing by a low priority task

Optimality and U_{lub}

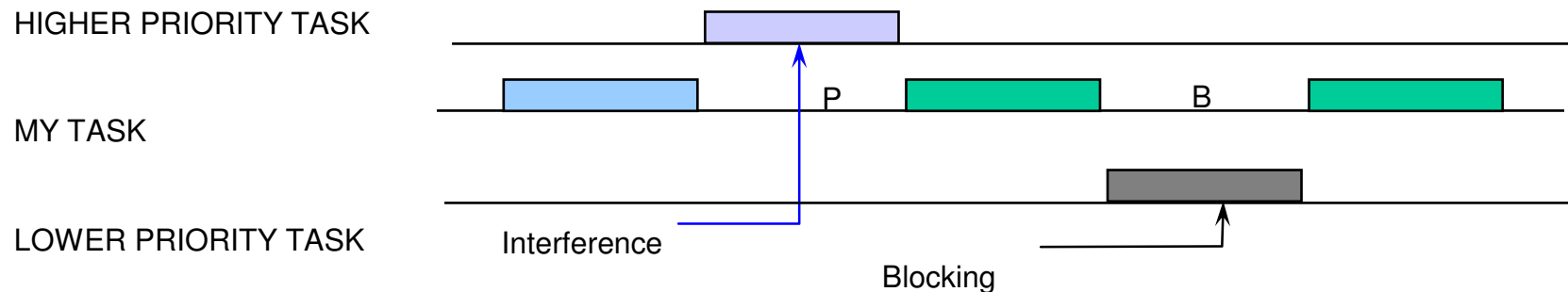
- When there are shared resources ...
 - The RM priority assignment is no more optimal. As a matter of fact, there is no optimal priority assignment (NP-complete problem [Mok])
 - The least upper bound on processor utilization can be arbitrarily low
 - It is possible (and quite easy as a matter of fact) to build a sample task set which is not schedulable in spite of a utilization $U \rightarrow 0$

Key concepts

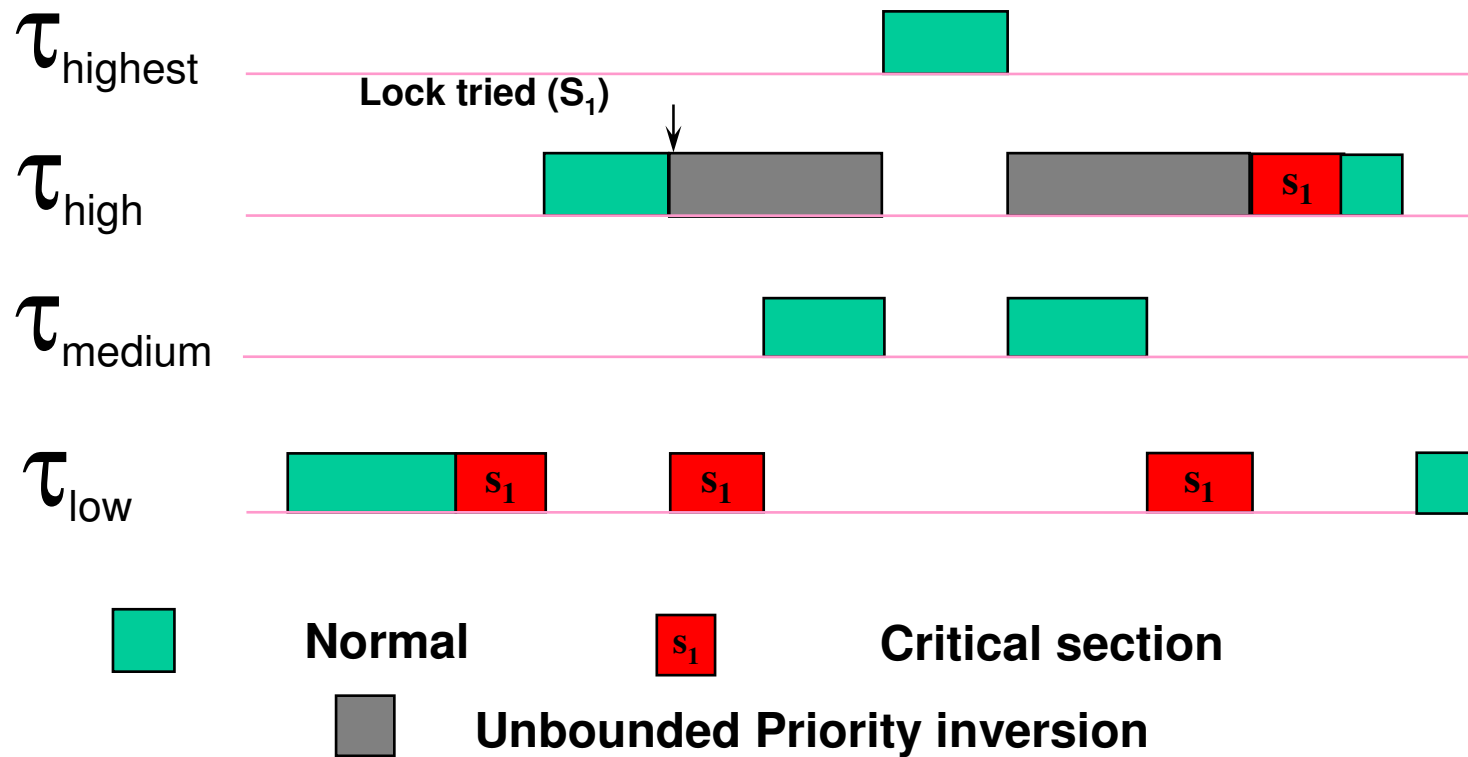
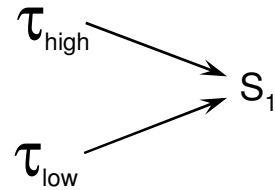
- Task
 - Encapsulating the execution thread
 - Scheduling unit
 - Each task implements an active object
- Protected Objects
 - Encapsulating shared information (Resources)
 - The execution of operations on protected objects is mutually exclusive

Response time of a real-time thread

- Execution time
 - time spent executing the task (alone)
- Execution of non schedulable entities
 - Interrupt Handlers
- Scheduling interference
 - Time spent executing higher priority jobs
- Blocking time
 - Time spent executing lower priority tasks
 - Because of shared resources



An example of “unbounded” priority inversion



Methods

- Non-preemptable CS
- Priority Inheritance
- Priority Ceiling (Original Priority Ceiling Protocol)
- Immediate priority ceiling or highest locker (Stack Resource Protocols)

Non-preemptable CS

- A task cannot be preempted if in critical section
- When a task enters a CS its priority is raised to the highest possible value

Advantages

- Simple and effective
- Prevents deadlocks
- Predictable!

Disadvantages

- May block tasks (even highest priority!) regardless of the fact that they use (some) resource or not ...
- Blocking term $B_i = \max(CS_{ip})$

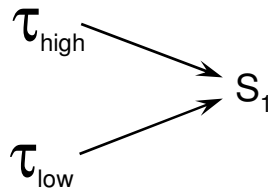
Preemption vs. non preemption

<i>Task</i>	C_i	T_i	π_i	<i>WCRT</i> (<i>P</i>)	<i>WCRT</i> (<i>NP</i>)	D_i^1	D_i^2
τ_1	20	70	1	20	20+35 = 55	45	60
τ_2	20	80	2	20 + 20 = 40	20+35+20 = 75	80	80
τ_3	35	200	3	35+2*20+2*20 = 115	35+20+20 = 75	120	100

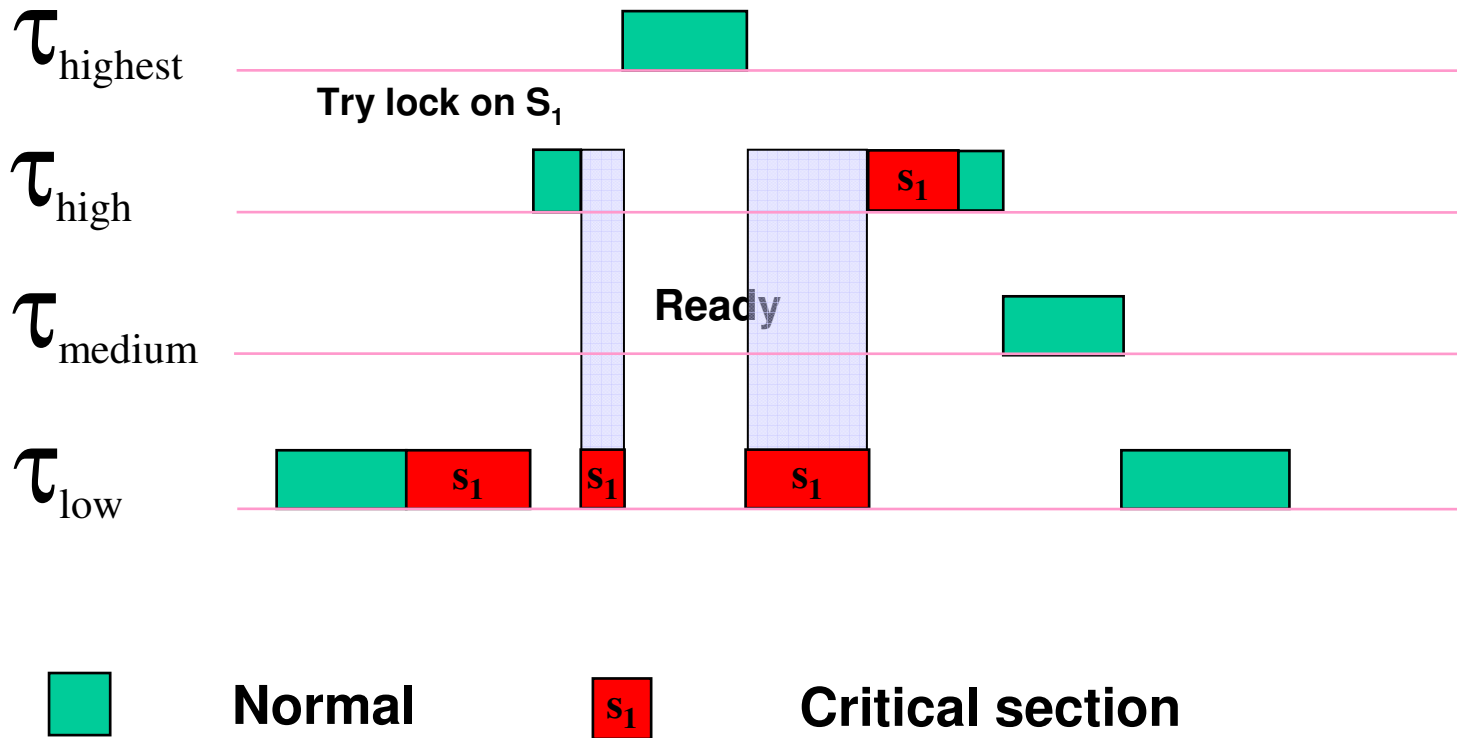
Priority Inheritance Protocol

- [Sha89]
- Tasks are only blocked when using CS
- Avoids unbounded blocking from medium priority tasks
- It is possible to bound the worst case blocking time if requests are not nested
- Saved the Mars Pathfinder ...

Priority Inheritance Protocol

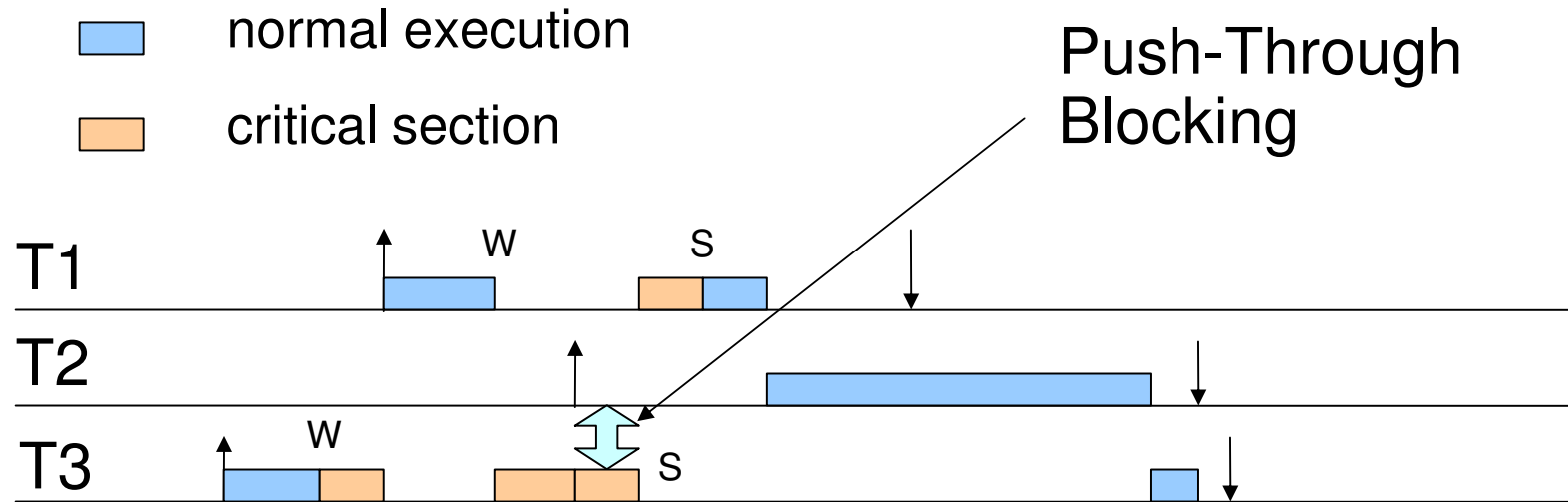


- High and low priority tasks share a common resource
- A task in a CS inherits the highest priority among all tasks blocked on the same resource

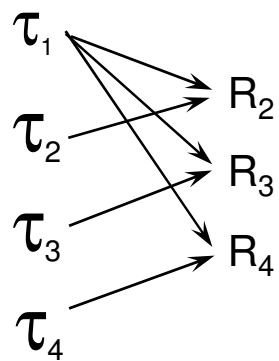


priority inheritance

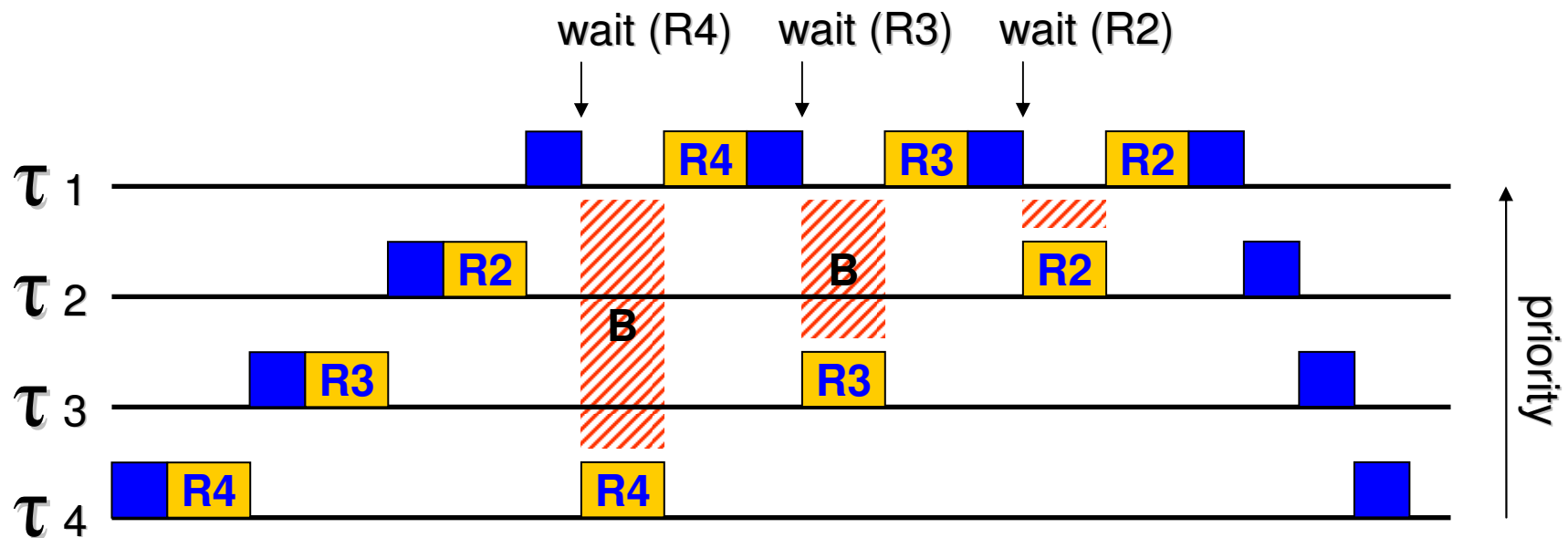
- low priority task inherits the priority of T1
- T2 is delayed because of push-through blocking (even if it does not use resources!)



Priority Inheritance Protocol: multiple blocking



Each task τ_i may block K times where
 $K = \min(nt_{lp(i)}, nr_{usage(i,k)})$



Priority Inheritance Protocol

- **Disadvantages**
- Tasks may block multiple times
- Worst case behavior (CS not nested) even worse than non-preemptable CS
- Costly implementation except for very simple cases
- Does not even prevent deadlock (nested CS)

Priority Ceiling Protocol

- *priority ceiling* of a resource S = maximum priority among all tasks that can possibly access S
- A process can only lock a resource if its dynamic priority is higher than the ceiling of any currently locked resource (excluding any that it has already locked itself).
- If task τ blocks, the task holding the lock on the blocking resource inherits its priority
- Two forms
 - Original ceiling priority protocol (OCPP)
 - Immediate ceiling priority protocol (ICPP, similar to Stack Resource Policy SRP)
- Properties (on single processor systems)
 - A high priority process can be blocked at most once during its execution by lower priority processes
 - Deadlocks are prevented
 - Transitive blocking is prevented

Deadlock (prevention)

- **Conditions for deadlock (Coffman 71)**

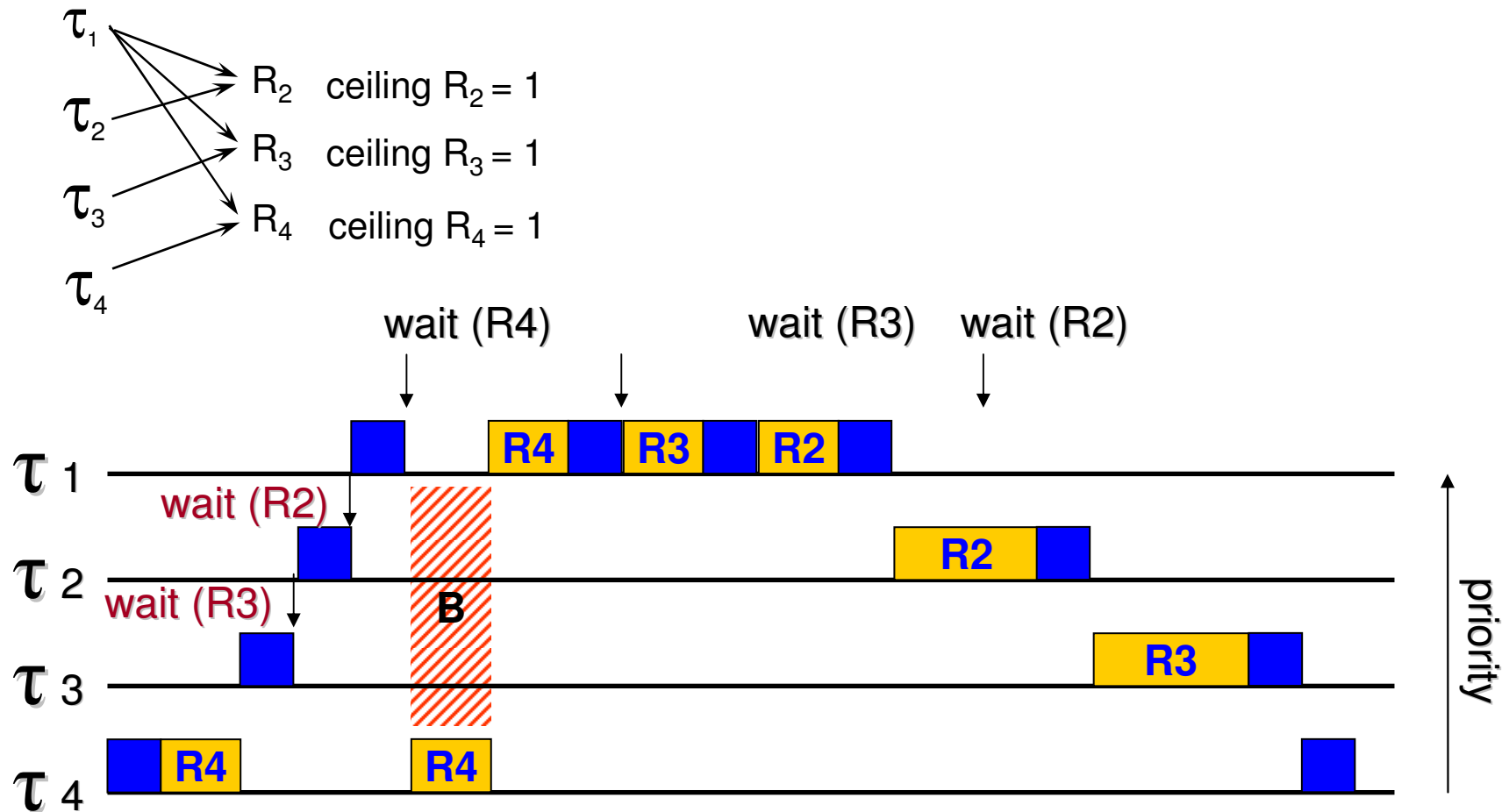
1. *Mutual exclusion* : a resource cannot be used by more than one process at a time
2. *Hold and wait* : processes already holding resources may request new resources
3. *No preemption*: No resource can be forcibly removed from a process holding it, Resources can be released only by the explicit action of the process

4. *Circular wait*: two or more processes form a circular chain where each process waits for a resource that the next process in the chain holds

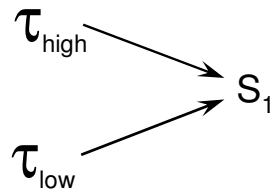
- Deadlock only occurs when all of the previous four hold true

PCP prevents circular waits!

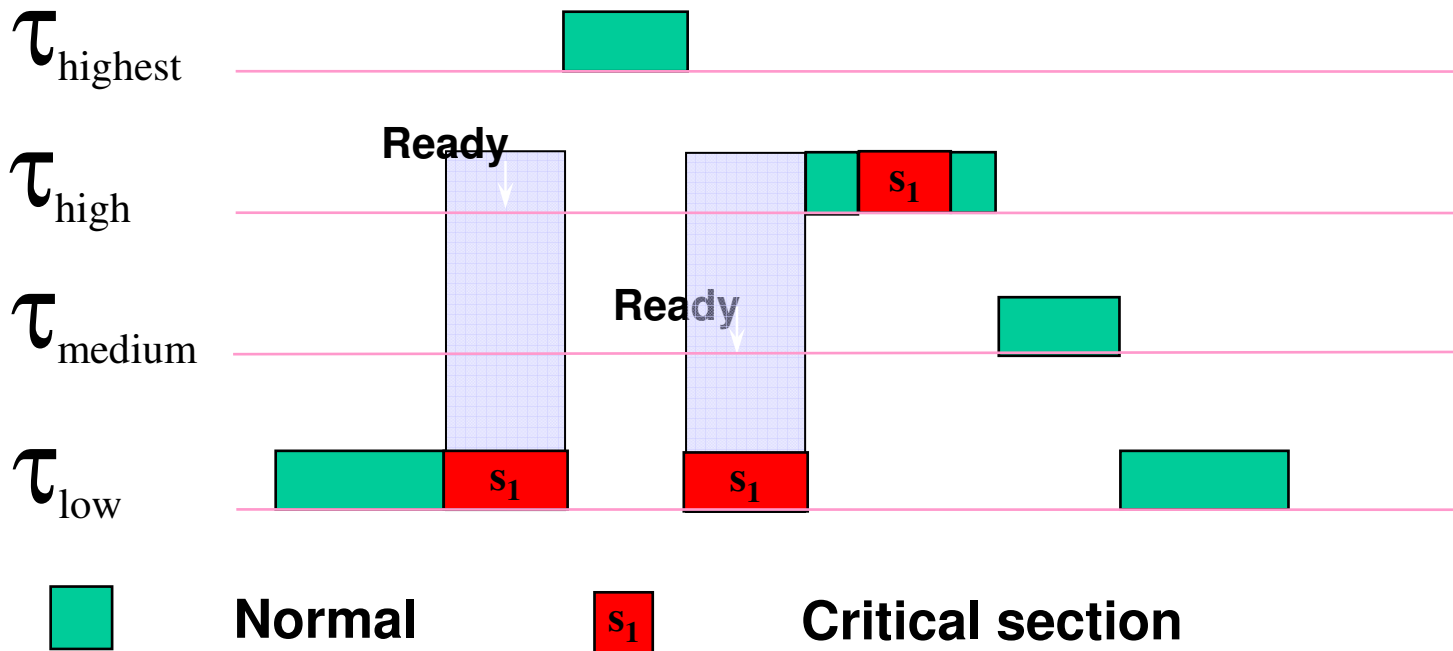
Example of OCPP



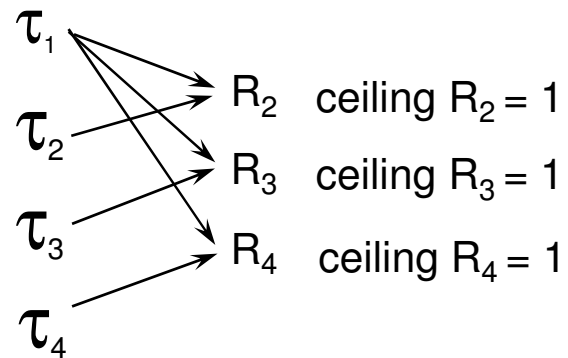
Immediate Priority Ceiling Protocol



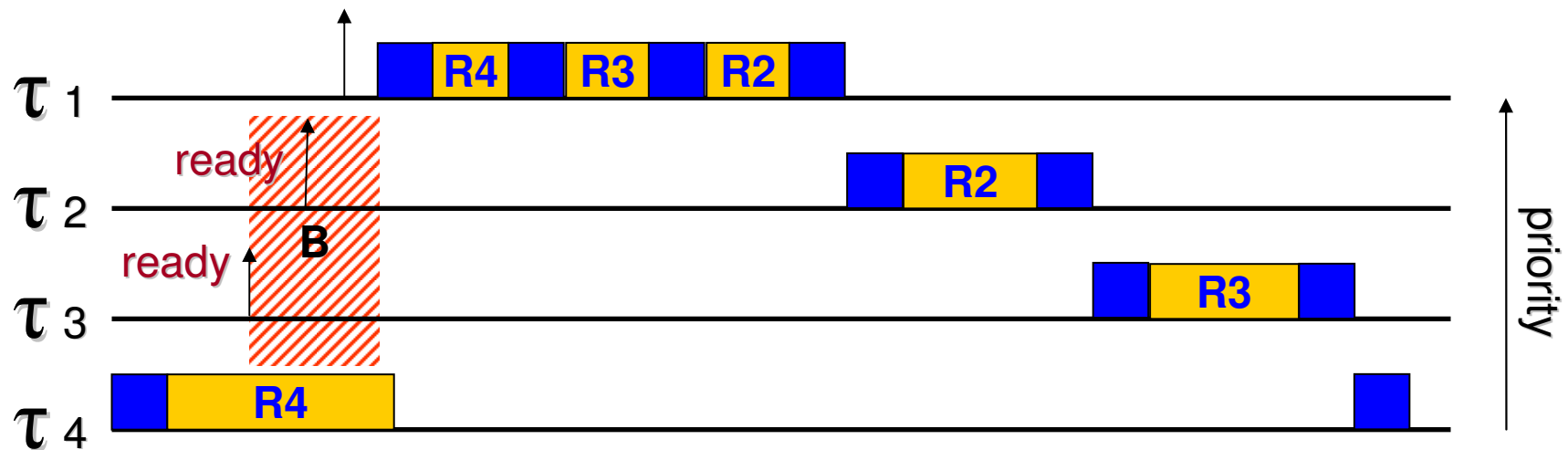
- High and low priority task share a critical section
- Ceiling priority of CS = Highest priority among all tasks that can use CS
- CS is executed at ceiling priority



Example of ICPP



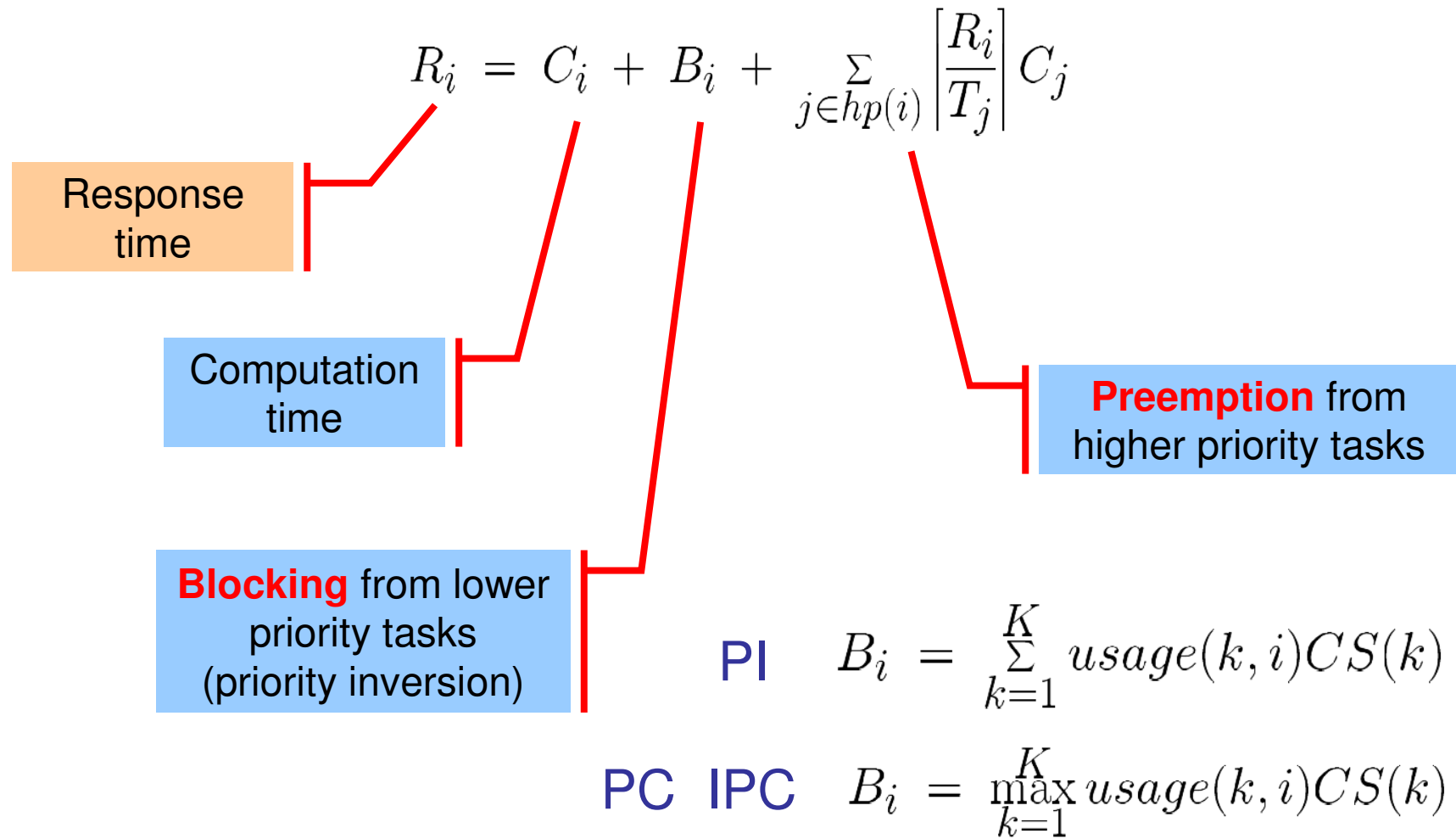
Execution of tasks is perfectly nested !



OCPP vs. ICPP

- Worst case behavior identical from a scheduling point of view
- ICPP is easier to implement than the original (OCPP) as blocking relationships need not be monitored
- ICPP leads to less context switches as blocking is prior to first execution
- ICPP requires more priority movements as this happens with all resource usages; OCPP only changes priority if an actual block has occurred.

Response time analysis



Response Time Calculations & Blocking (contd.)

$$\text{PI} \quad B_i = \sum_{k=1}^K \text{usage}(k, i) CS(k)$$

$$\text{PC IPC} \quad B_i = \max_{k=1}^K \text{usage}(k, i) CS(k)$$

- Where usage is a 0/1 function:

$$\text{usage}(k, i) = 1$$

if resource k is used by at least one

process with a priority less than i, and at

least one process with a priority greater or equal to i.

Otherwise it gives the result 0.

- CS(k) is the computational cost of executing the longest k-th critical section called by a lower priority task .

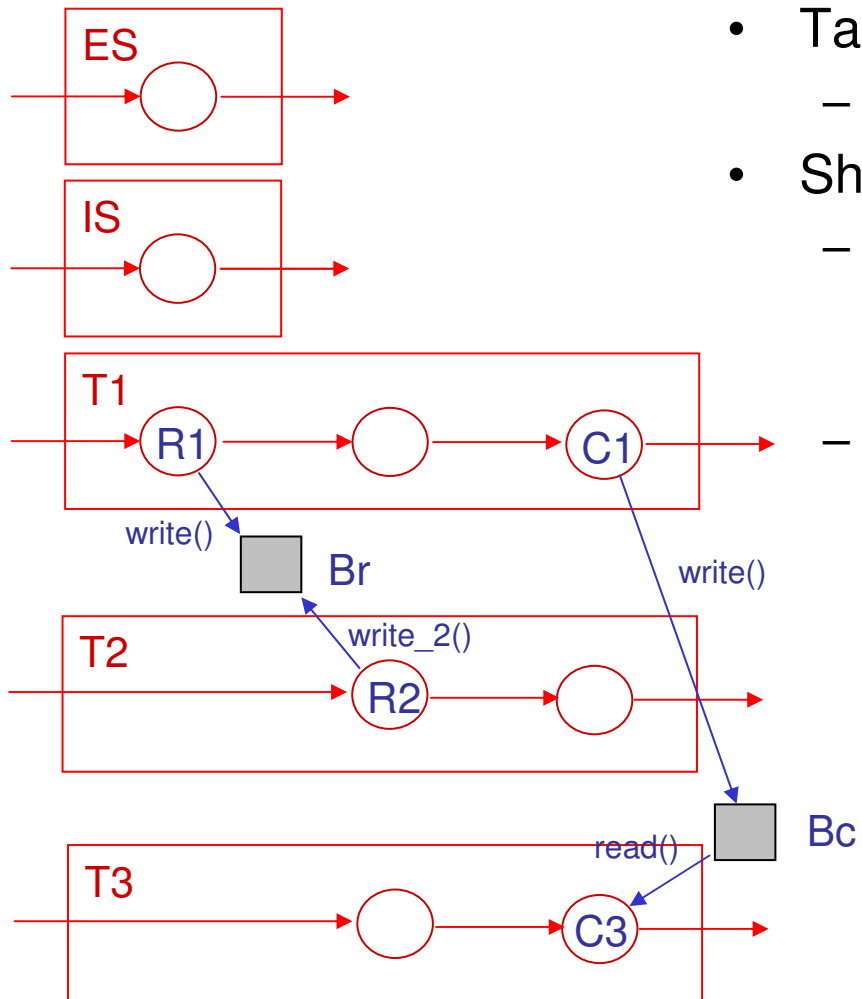
Blocking time in PCP and IPCP

- An example ...

	R1	R2	R3	B_{PIP}	B_{PCP}
$\tau 1$		20		5	5
$\tau 2$	5		10	20	10
$\tau 3$		5	5	18	10
$\tau 4$			5	13	10
$\tau 5$	10	3			

The diagram illustrates the blocking time in PCP and IPCP. The table shows the values for tasks $\tau 1$ to $\tau 5$ across resources R1, R2, R3, and blocking times B_{PIP} and B_{PCP} . Red arrows indicate dependencies between tasks and resources. Yellow circles highlight specific values in the R1, R2, and R3 columns.

Example: Shared resources



- Task
 - 5 Tasks
- Shared resources
 - Results buffer
 - Used by R1 and R2
 - R1 (2 ms) R2 (20 ms)
 - Communication buffer
 - Used by C1 and C3
 - C1 (10 ms) C3 (10 ms)

Example: Shared Resources

<i>Task</i>	<i>C</i>	<i>T</i>	π	<i>WCRT</i> (<i>PI</i>)	<i>WCRT</i> (<i>PC</i>)	<i>D</i>
<i>ES</i>	5	50	1	5+0+0 =5	5+0+0 =5	6
<i>IS</i>	10	100	2	10+0+5 = 15	10+0+5 =15	100
<i>T1</i>	20	100	3	20+30+20 = 70	20+20+20 =60	100
<i>T2</i>	40	150	4	40+10+40 =90	40+10+40 =90	130
<i>T3</i>	100	350	5	100+0+200 =300	100+0+200 =300	350

Blocking factor in the sufficient schedulability formula

- Let B_i be the duration in which τ_i is blocked by lower priority tasks
- The effect of this blocking can be modeled as if τ_i 's utilization were increased by an amount B_i/T_i
- The effect of having a deadline D_i before the end of the period T_i can also be modeled as if the task were blocked for $E_i=(T_i-D_i)$ by lower priority tasks
 - as if utilization increased by E_i/T_i

Scheduling with Offsets

- Enhanced model ...
- Each periodic task τ_i is characterized by the quadruple (T_i, D_i, C_i, O_i)
- The offset O_i is the instant of the first request
- The requests of τ_i are separated by T_i time units and occur at time $O_i + (k - 1)T_i$ ($k=1, 2, \dots$).
- The execution of the k -th request of task τ_i , which occurs at time $O_i + (k - 1)T_i$, must finish before or at time $O_i + (k - 1)T_i + D_i$

Scheduling with Offsets

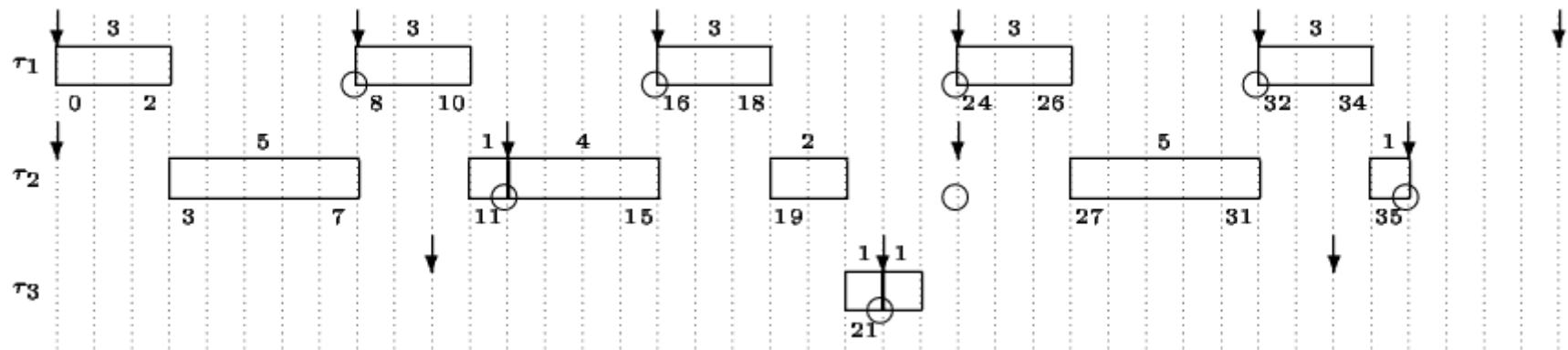
- **Synchronous model:** all $O_i=0$
- **Asynchronous model:** O_i may be $\neq 0$, but the values are given
- **Offset free model:** the values of the O_i may be defined to improve schedulability

Scheduling with Offsets

- The synchronous case is the worst case, hence it is clearly pessimistic ...
 - Example:
 - consider the case (C, D, T)
- $\tau_1=(3, 8, 8)$, $\tau_2=(6, 12, 12)$, $\tau_3=(1, 12, 12)$
- The example is not schedulable in the synchronous case (not even with RM priority assignment), given that task 3 misses its deadline.

Scheduling with Offsets

- But if you try $O_1=0$, $O_2=0$, $O_3=10$ the set becomes schedulable !



Scheduling with Offsets

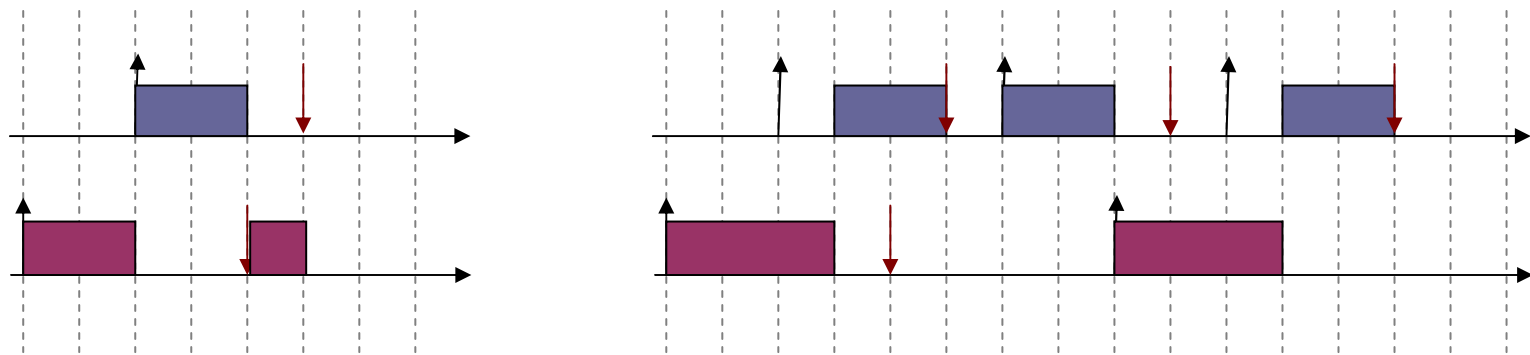
- This brings the promise for an increase in schedulability
 - Unless all periods are prime, in which case there is still a critical instant!
- Unfortunately ...
- The RM/DM priority assignments are no more optimal!
 - We need a way to find a (possibly optimal) priority assignment
- There is no critical instant
 - the standard response time test is not valid anymore
 - We need a (possibly efficient) schedulability test
- There is availability of a priority assignment method and a schedulability test !
 - but what is really needed is an offset synthesis procedure !

Scheduling with Offsets

- The (RM/DM) priority assignment is (in general) not optimal.
- counterexample (O, C, D, T)

$$\tau_1=(2, 2, 3, 4), \quad \tau_2=(0, 3, 4, 8)$$

- if τ_1 has priority higher than τ_2 (as in RM/DM) deadline is missed at time 4.
- When priorities are reversed, deadlines are met ...



Scheduling with Offsets

- Feasibility test for asynchronous task sets:
- Given a set of offsets O_1, O_2, \dots, O_n
- [Leung82] a task set is feasible if all deadlines are met in $[s, 2P]$, where $s = \max\{O_1, O_2, \dots, O_n\}$ and $P = \text{lcm}\{T_1, T_2, \dots, T_n\}$
 - in practice it is sufficient to build the schedule and check all the busy periods originating from a task release time in $[s, 2P)$
- For fixed priority tasks and $D \leq T$ it is possible to further restrict the interval [Audsley91]

THEOREM 7 *Let S_i be inductively defined by $S_1 = O_1$, $S_i = \max\{O_i, O_i + \lceil \frac{S_{i-1} - O_i}{T_i} \rceil T_i\}$ ($i = 2, 3, \dots, n$), then if the task set is ordered by decreasing priorities and has a feasible schedule, it is periodic from S_n with period $P = \text{lcm}\{T_i | i = 1, \dots, n\}$.*

- This means that it is sufficient to check the interval $[S_n, S_n + P]$

Scheduling with Offsets

- Now we do have a test and the algorithm by audsley to find an optimal fixed priority assignment for the case $D > T$ works in this case as well ...
- But ...
- the most important case (offset free systems) requires a procedure for setting up the offsets.
- The problem of finding the optimal offset assignment is probably NP-complete [Goosens00]
- Approximate solutions are sought ...

Scheduling with Offsets

- An example: dissimilar offset assignment [Goossens00]

Algorithm 1 The dissimilar offset assignment

```
1:  $G \leftarrow \{(i, j, \gcd(T_i, T_j)) \mid 1 \leq i < j \leq n\}$ ;  
2:  $\mathcal{G} \leftarrow ((i_1, j_1, \gcd(T_{i_1}, T_{j_1})), (i_2, j_2, \gcd(T_{i_2}, T_{j_2})), \dots)$ ,  
   with  $\{(i_k, j_k, \gcd(T_{i_k}, T_{j-k})) \mid k = 1, \dots, \frac{n(n-1)}{2}\} = G$ , such that  $r < p \implies$   
    $\gcd(T_{i_r}, T_{j_r}) \geq \gcd(T_{i_p}, T_{j_p})$ ;  
3: {The vector  $\mathcal{G}$  is a sorted version of the set  $G$ . In the following we shall use the  
   "dot notation" to denote the 3 fields of each entry of  $\mathcal{G}$ , which are row, col and  
   gcd, respectively.}  
4: assignment  $\leftarrow n$ ; {The remaining number of offset assignments.}  
5: Mark  $\leftarrow$  (false,  $\dots$ , false); { $n$  components.}  
6:  $k \leftarrow 1$ ;  
7: while assignment  $> 0$  do  
8:   if  $\neg(\text{Mark}_{\mathcal{G}_k.col}) \wedge \neg(\text{Mark}_{\mathcal{G}_k.row})$  then  
9:      $O_{\mathcal{G}_k.row} \leftarrow \text{rand}()$ ;  $O_{\mathcal{G}_k.col} \leftarrow O_{\mathcal{G}_k.row} + \mathcal{G}_k.gcd \text{ div } 2$ ;  
10:    assignment  $\leftarrow$  assignment  $- 2$ ;  $\text{Mark}_{\mathcal{G}_k.row} = \text{true}$ ;  $\text{Mark}_{\mathcal{G}_k.col} = \text{true}$ ;  
11:   else if  $\neg(\text{Mark}_{\mathcal{G}_k.col})$  then  
12:      $O_{\mathcal{G}_k.col} \leftarrow O_{\mathcal{G}_k.row} + \mathcal{G}_k.gcd \text{ div } 2$ ;  
13:     assignment  $\leftarrow$  assignment  $- 1$ ;  $\text{Mark}_{\mathcal{G}_k.col} = \text{true}$ ;  
14:   else if  $\neg(\text{Mark}_{\mathcal{G}_k.row})$  then  
15:      $O_{\mathcal{G}_k.row} \leftarrow O_{\mathcal{G}_k.col} + \mathcal{G}_k.gcd \text{ div } 2$ ;  
16:     assignment  $\leftarrow$  assignment  $- 1$ ;  $\text{Mark}_{\mathcal{G}_k.row} = \text{true}$ ;  
17:   end if  
18:    $k \leftarrow k + 1$ ;  
19: end while
```

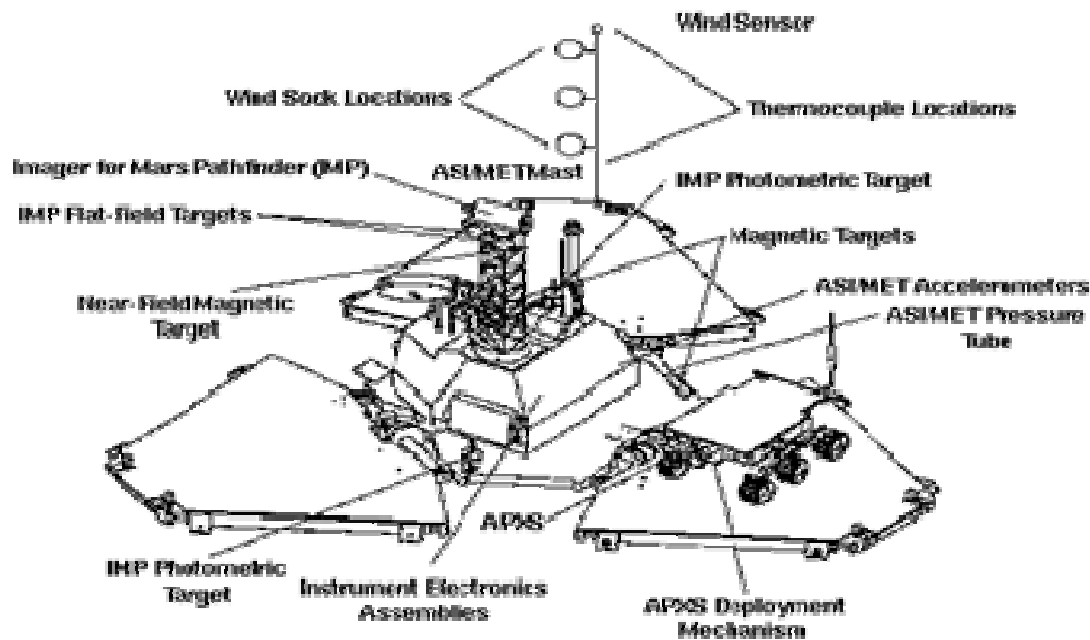
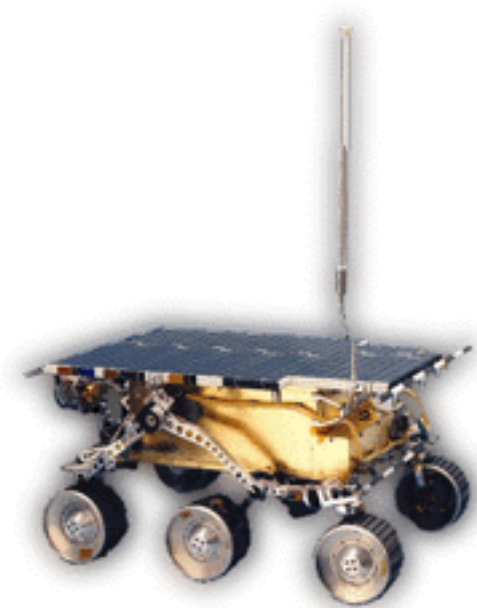

The MARS PATHFINDER

A Priority Inversion case



The Mars Pathfinder Case

- Overview
- MARS PATHFINDER – ARCHITECTURE
- THE 1553 BUS
- THE PROBLEM
- A PRIORITY INHERITANCE SOLUTION



Mars Pathfinder was the second mission in the NASA Discovery program.

Mission started on November 16th 1996 and finished on September 27th 1997.

The Mars Pathfinder Case

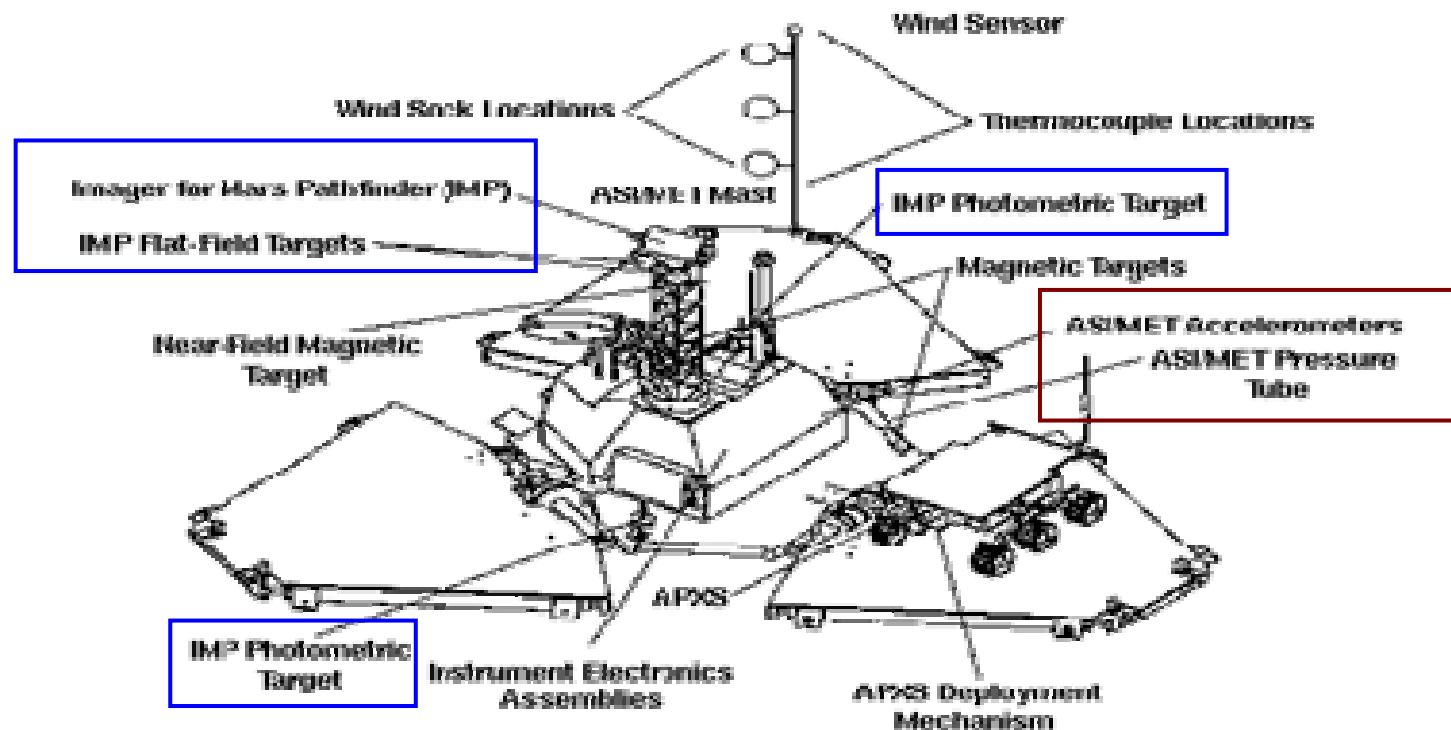
The system consists of two units:

cruiser / lander (fixed) hosting the navigation and landing functionality and the subsystems :

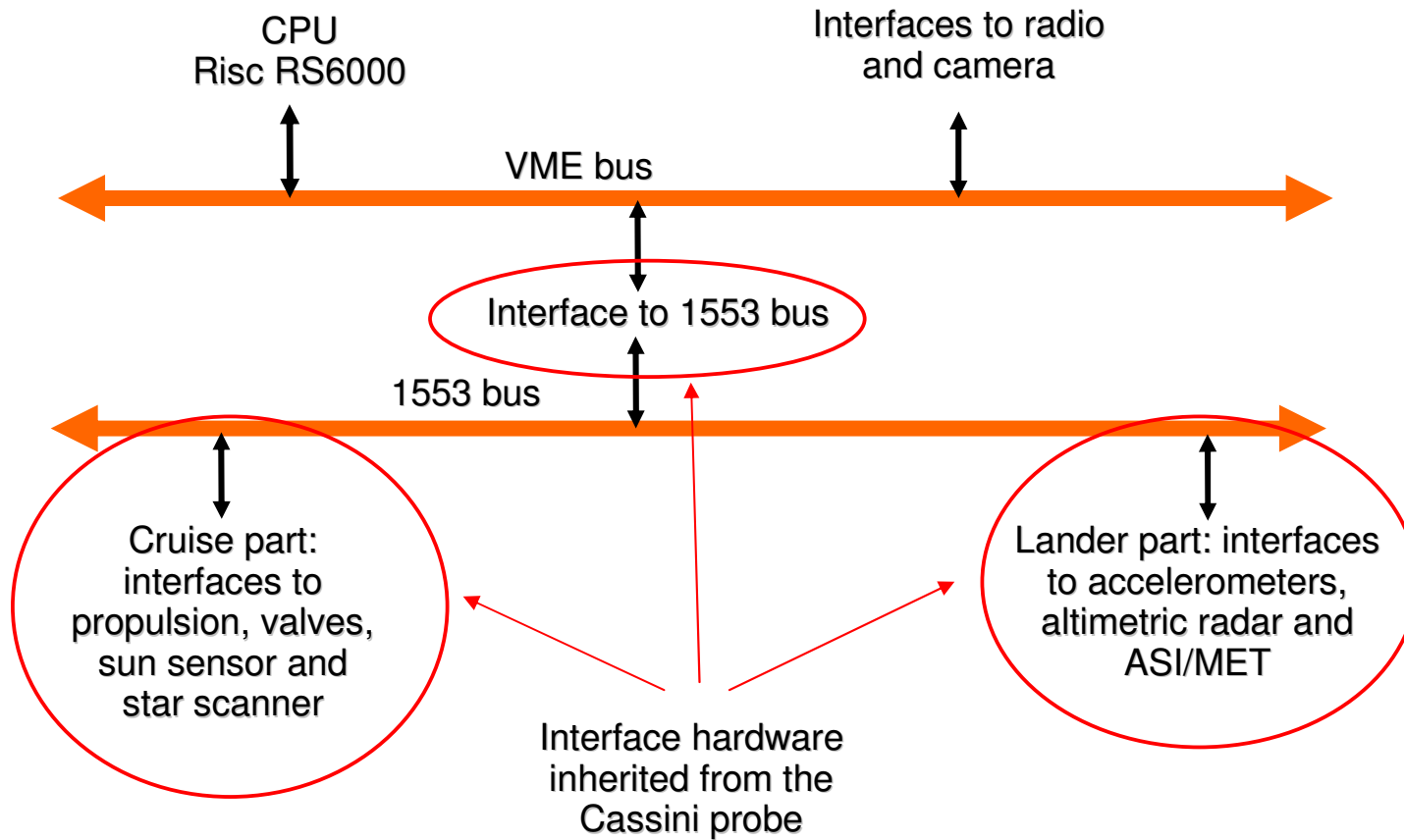
- **ASI/MET** for sensing meteorological atmospheric data
- **IMP** for image acquisition

microrover (mobile) hosting :

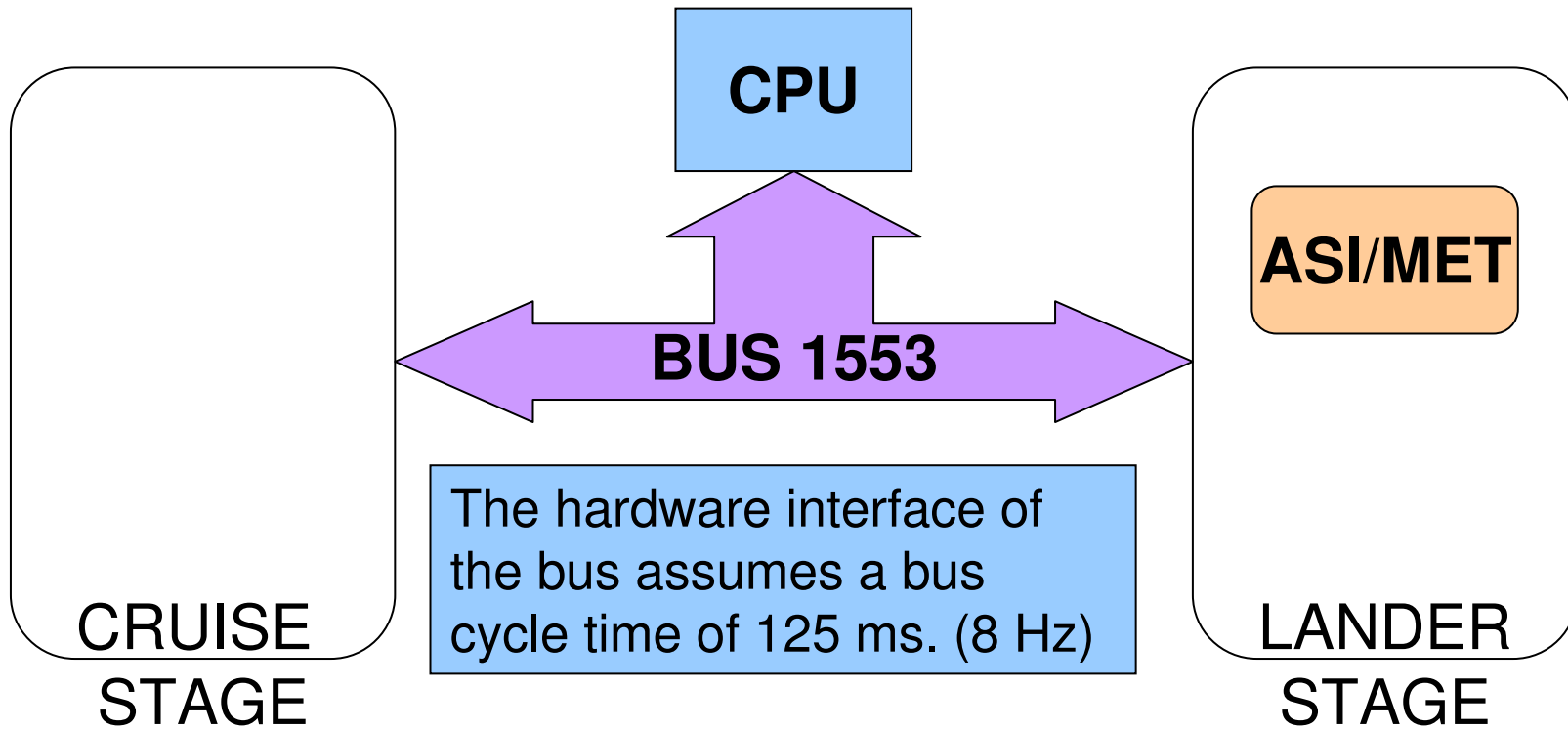
- **APXS** : X-ray spectrometer
- Image acquisition devices



System architecture



Architecture



Software Architecture

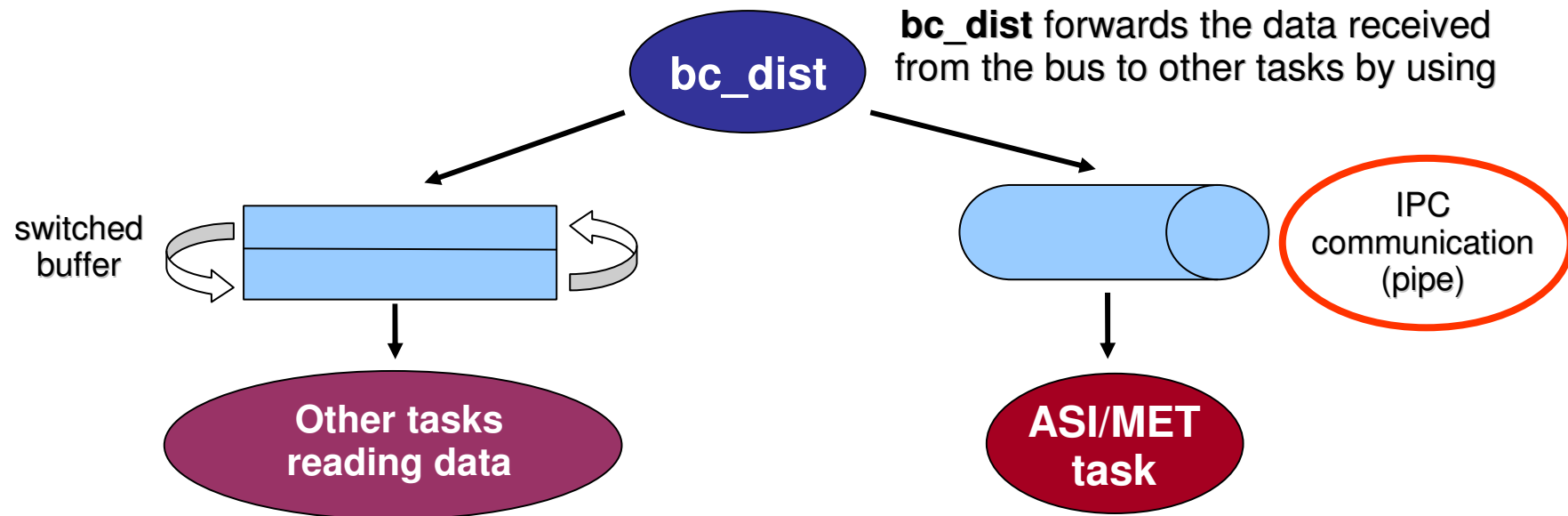
- Cyclic Scheduler @ 8 Hz
- The 1553 is controlled by two tasks:
 - Bus Scheduler: **bc_sched** computes the bus schedule for the next cycle by planning transactions on the bus (**highest priority**)
 - Bus distribution: **bc_dist** collects the data transmitted on the bus and distributes them to the interested parties (**third priority level**)
 - A task controlling entry and landing is second level, there are other tasks and idle time
- **bc_sched** must complete before the end of the cycle to setup the transmission sequence for the upcoming cycle.
 - In reality bc_sched and bc_dist must not overlap



What happened

- The Mars Pathfinder probe lands on Mars on July 4th 1997
- After a few days the probe experiences continuous system resets as a result of a detected critical (timing) error

Software architecture of the Pathfinder



IPC (InterProcess Communication mechanism)

- VxWorks provided POSIX pipes
- Files descriptors associated to the reading and writing sides of the pipe are shared resources protected by mutexes
- ASI/MET called a `select()` for reading data from the pipe

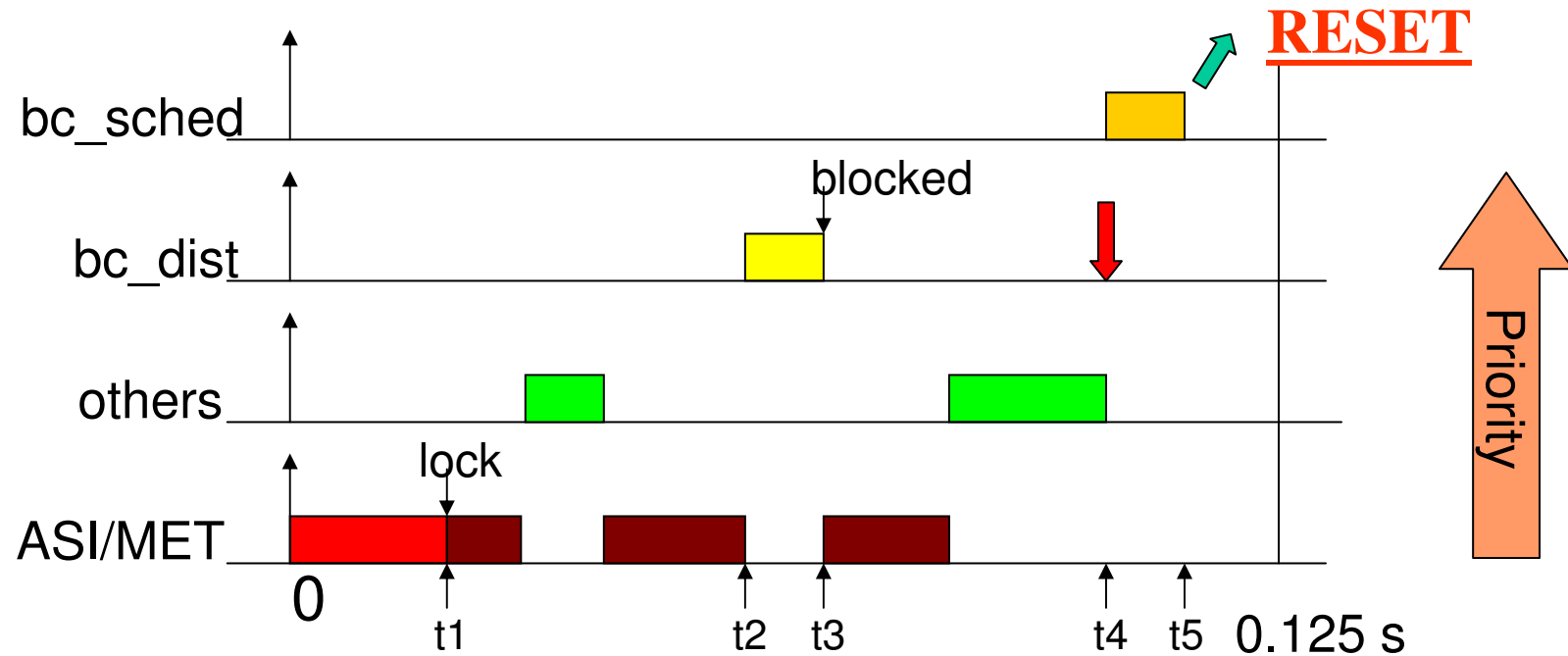
The problem

- The task responsible for system malfunctions is **ASI/MET**
- The ASI/MET task handles meteo data and transmits them using an IPC mechanism based on ***pipe()***
- Other tasks read from the pipe using the ***select()*** primitive, hiding a mutex semaphore
- Tasks in the system
 - bc_sched*** maximum priority
 - bc_dist*** priority 3
 - several medium priority tasks***
 - ASI/MET with low priority***
- ASI/MET calls `select()` but, before releasing the mutex, is preempted by medium priority tasks. `bc_dist`, when ready, tries to lock the semaphore that controls access to the pipe. The resource is taken by ASI/MET and the task blocks
- When `bc_sched` starts for setting the new cycle, it detects that the previous cycle was not completed and resets the system.

The problem

- The select mechanism creates a mutual exclusion semaphore to protect the "wait list" of file descriptors
- The ASI/MET task had called select, which had called pipelock(), which had called selNodeAdd(), which was in the process of giving the mutex semaphore. The ASI/ MET task was preempted and semGive() was not completed.
- Several medium priority tasks ran until the bc_dist task was activated. The bc_dist task attempted to send the newest ASI/MET data via the IPC mechanism which called pipeWrite(). pipeWrite() blocked, taking the mutex semaphore. More of the medium priority tasks ran, still not allowing the ASI/MET task to run, until the bc_sched task was awakened.
- At that point, the bc_sched task determined that the bc_dist task had not completed its cycle (a hard deadline in the system) and declared the error that initiated the reset.

The priority inversion

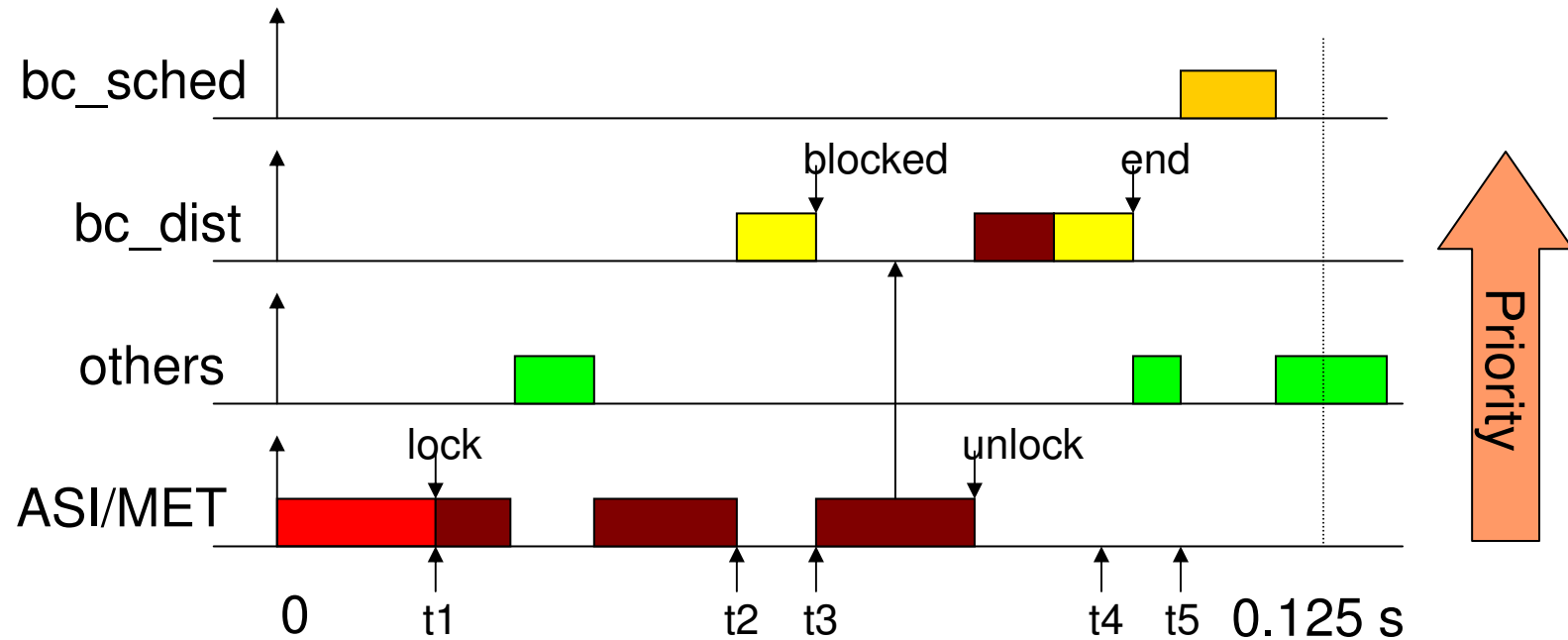


- `ASI/MET` acquires control of the bus (shared resource)
- Preemption of `bc_dist`
- Lock attempted on the resource
- `bc_sched` is activated, `bc_dist` is in execution after the deadline
- `bc_sched` detects the timing error of `bc_dist` and resets the system

The Solution

- After debugging on the pathfinder replica at JPL, engineers discover the cause of malfunctioning as a **priority inversion problem**.
- Priority Inheritance was disabled on pipe semaphores
- The problem did not show up during testing, since the schedule was never tested using the final version of the software (where medium priority tasks had higher load)
- The on-board software was updated from earth and semaphore parameters (global variables in the `selectLib()`) were changed
- The system was tested for possible consequences on system performance or other possible anomalies but everything was OK

Pathfinder with PIP



ASI/MET is not interrupted by medium priority tasks since inherits bc_dist priority.

Should you use PIP?

- See “Against priority Inversion” [Yodaiken] available from the web
- Critical sections protected by PIP semaphores produce a large worst case blocking term
 - chain-blocking
 - The blocking factor is the **sum** of the worst case length of the critical sections (plus protocol overhead)
- PIP does not support nested CS with bounded blocking (very difficult to guess where implementation of OS primitives such as pipe operations implies CS)

Should you use PIP?

- Except for very simple (but long) CS, PIP does not provide performances better than other solutions (non preemptive CS or PCP)
- PIP has a costly implementation, overheads include:
 - Managing the basic priority inheritance mechanism not only requires updating the priority of the task in CS, but handling a complex data structure (not simply a stack) for storing the set of priorities inherited by each task (one list for each task and one for each mutex)
 - Dynamic priority management implies dynamic reordering of the task lists
- For a full account ...

http://research.microsoft.com/~mbj/Mars_Pathfinder/Authoritative_Account.html

That's all folks !

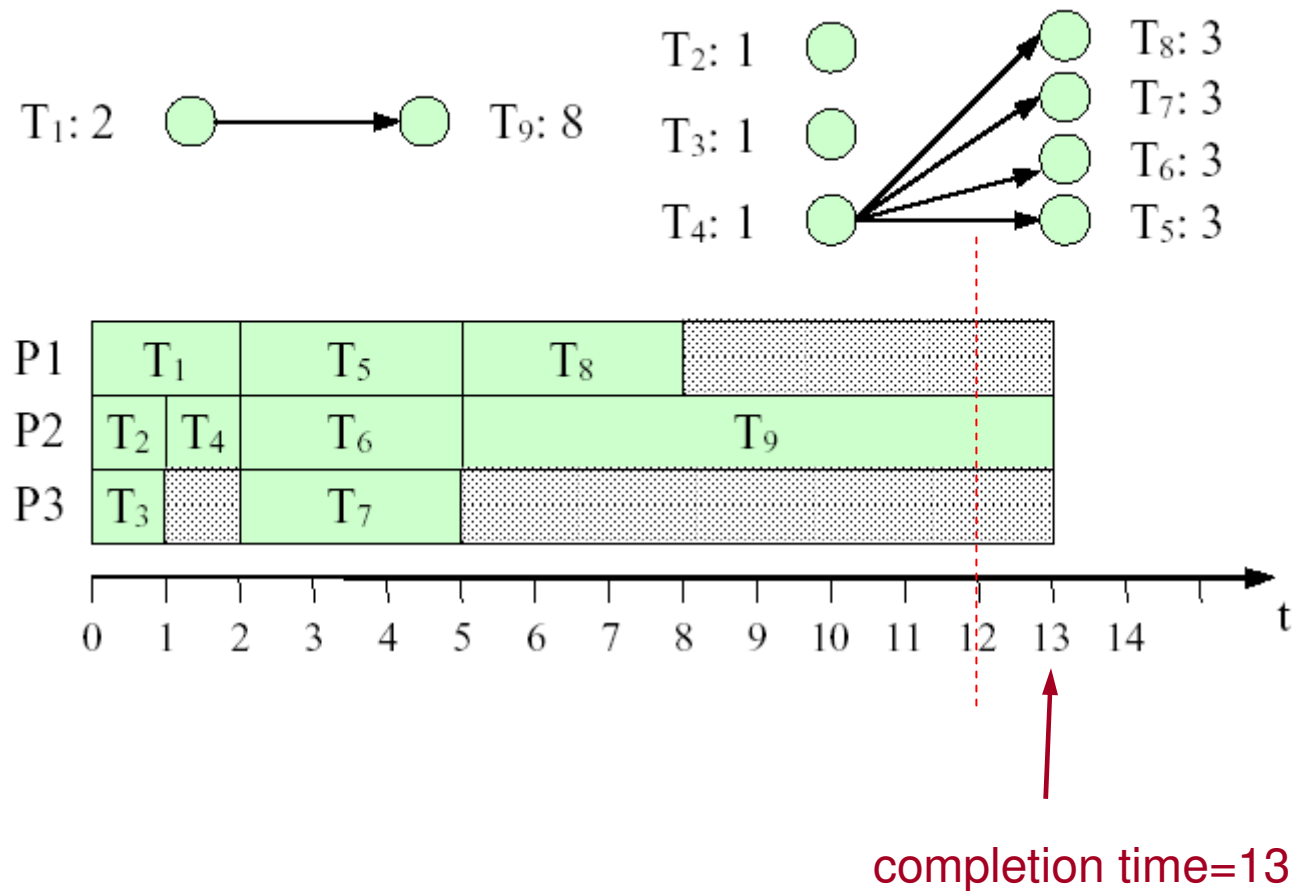
- Please ask your questions
...



- Backup slides 1- Anomalies and cyclic schedulers

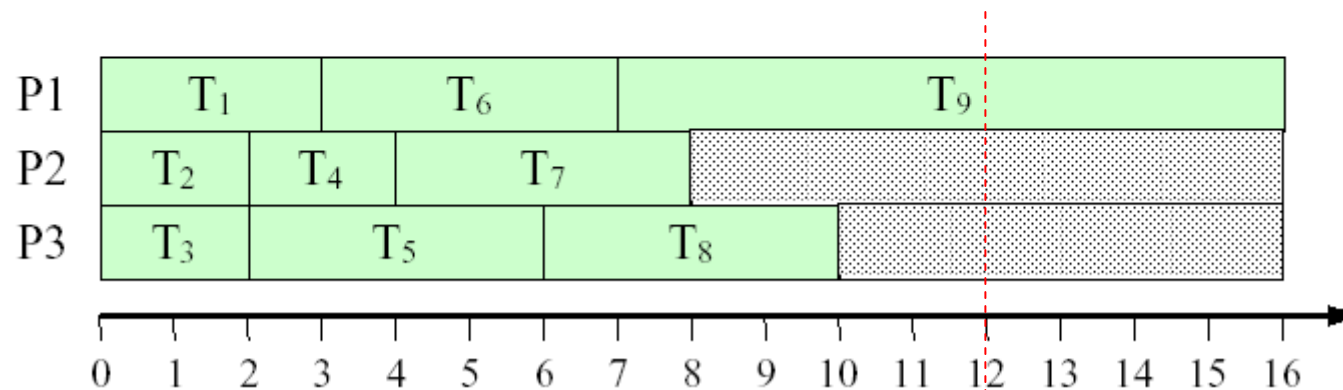
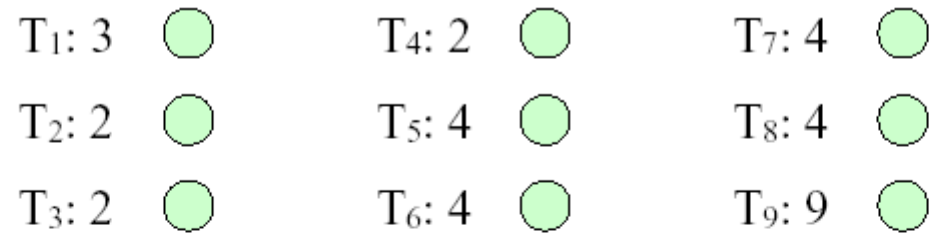
Operating Systems background

- Shortening tasks



Operating Systems background

- Releasing precedence constraints



completion time=16